Extended Decorrelation Procedures for Source Separation of Biomedical Image Data

Holger Schöner
hfsch@cs.tu-berlin.de

Diploma Thesis

Technical University of Berlin
Department of Computer Science
Neural Information Processing Group
May 1999

Supervisors:
Prof. Dr. K. Obermayer
Dr. M. Stetter

Contents

Chapter 1

Übersicht

Die Funktionsweise des menschlichen Gehirns beschäftigt Philosophen und Naturwissenschaftler seit langer Zeit. Es zeigt sich, daß Aufgaben, die für das menschliche Gehirn einfach zu sein scheinen, wie zum Beispiel die Erkennung von Gesichtern, nur sehr schwer oder gar nicht von heutigen Maschinen oder Computern erledigt werden können. Andererseits haben letztere Fähigkeiten, mit denen das menschliche Gehirn nur wesentlich schlechter zurecht kommt, wie mathematische Rechnungen oder Informationsspeicherung. Wenn komplexe Vorgänge im Gehirn, z.B. Gesichtererkennung, Sprachverständnis oder Abstraktion und Schließen, besser verstanden würden, könnte dies viele neue Möglichkeiten eröffnen.

Eine Art, ein solches Verständnis zu erreichen, ist die Untersuchung der Informationsverarbeitung und -speicherung im Gehirn. Neuronale Karten, die die Aktivierung verschiedener Teile des Cortex für unterschiedliche Stimuli oder während der Ausführung von Aktionen zeigen, sind für diesen Zweck sehr hilfreich. Die Aufzeichnungen von Optical-Imaging Experimenten, wie sie auch in dieser Arbeit untersucht werden, können dafür verwendet werden, solche Karten zu erstellen. In solchen Aufzeichnungen werden zum einen Signale aufgezeichnet, die mit der neuronalen Aktivität verknüpft sind; andererseits aber auch störende Signale, wie Adern, mit dem Puls zusammenhängende Oszillationen, etc.

Die übliche Art, aus diesen Aufzeichnungen die aktivitätsbezogenen Komponenten herauszufiltern, ist die Benutzung von Bandpaßfiltern, in Verbindung mit Verfahren, die das Signal-Rausch-Verhältnis verbessern, z.B. das Aufsummieren mehrerer Experimente. Bandpaßfilter sind jedoch problematisch, da durch ihre Verwendung die Karten so verändert werden können, daß wichtige Statistiken, z.B. die Anzahl von Singularitäten in Orientierungspräferenzkarten, nicht mehr stimmen.

Ein anderer Ansatz, der in letzter Zeit zur Gewinnung von neuronalen Karten benutzt wird, ist Blind Source Separation (BSS). Dieser versucht, das aktivitätsbezogene Signal durch lineares Entmischen von den anderen Signalen zu trennen. Dabei existieren verschiedene Verfahren, die Entmischung zu lernen. Das hier näher betrachtete Verfahren, der Extended Spatial Decorrelation (ESD) Ansatz ([MS94, SSM$^+$99]), lernt die Entmischungsmatrix nur aus den beobachteten Daten, unter Verwendung derer (verschobener) Korrelationen.

Einige der zur Verfügung stehenden BSS-Verfahren werden in [SSM$^+$99] auf ihre Eignung für Optical-Imaging Experimente untersucht. Obwohl sie grundsätzlich in der Lage sind, Karten zu extrahieren, stellt sich ihre Anfälligkeit für Sensor-Noise (Rauschen, das nach dem Mischen

auftritt, z.B. Kamerarauschen) als Problem dar. Einer der in dem Artikel getesteten Algorithmen, die ESD Methode, wird in dieser Diplomarbeit bezüglich der Rauschrobustheit weiterentwickelt und mit zwei Datensätzen getestet.

Die Weiterentwicklung besteht im Wesentlichen darin, statt einer Verschiebung mehrere für die Berechnung von Kreuz-Korrelationen heranzuziehen. Dies vermindert den Einfluß von Sensor-Noise und macht den Algorithmus weniger sensitiv bezüglich der Wahl der einzelnen Verschiebung. Als Optimierungsmethode für die resultierende erweiterte Fehlerfunktion wird ein beschleunigter Gradientenabstieg benutzt. Dieser ist zwar abhängig von der Initialisierung seiner Parameter, dafür aber flexibler bezüglich der gefundenen Entmischungsmatrizen, verglichen mit anderen Multi-Shift Algorithmen.

Um die Rauschrobustheit beurteilen zu können, wird ein künstlicher Datensatz benutzt, für den der Rauschanteil kontrolliert werden kann. Verglichen werden die Leistungsfähigkeit von Single- und Multi-Shift-Algorithmen, das Verhalten für übliches Sphering (eine Vorverarbeitung der Daten, die für einige Algorithmen nötig, für andere hilfreich ist) und für rausch-robustes Sphering, und der Einfluß von räumlich korreliertem Rauschen im Gegensatz zu weißem Rauschen. Das Ergebnis zeigt eine im Vergleich zu den anderen Algorithmen größere Rauschrobustheit des hier entwickelten Algorithmus'. Simulationen mit dem auch in [SSM$^+$99] verwendeten Augendominanz-Datensatz zeigen für den neuentwickelten Algorithmus vorteilhafte Resultate; die extrahierten Karten haben eine sehr gute Qualität. Die Trennung des Mapping-Signals von Artefakten wie Blutgefäßen, globalem Signal, etc. ist deutlich besser als bei herkömmlichen Verfahren.

Chapter 2

Scope of Thesis

The way the human brain works has fascinated philosophers and scientists since a long time. It is hard to imitate tasks, which seem to be easy for the human brain, like visually recognizing faces, with machines or computers. On the other hand, these have capabilities the brain is not very powerful in, like doing calculations and storing information. It would open up many new possibilities, if processes like face recognition, language understanding, abstraction and inference in the human brain were comprehended.

One method applied to reach this comprehension is the analysis of how information is processed and represented in different parts of the brain. Maps of the activity of neurons in the cortex, for different stimuli or during certain actions are performed, are very useful for this method. The optical imaging experiments, which are examined in this thesis, have as a goal the extraction of such maps. Different signals indicating neural activity are recorded, together with unrelated signals like blood vessels, biological and recording noise, by these experiments.

Conventional optical imaging mostly uses, among other methods to improve the signal to noise ratio, bandpass filters to extract the activity maps. The use of bandpass filters is problematic, because the resulting maps and the statistics of their features (e.g. number of singularities in orientation preference maps) can be influenced by this.

A different approach recently used is the use of Blind Source Separation (BSS) methods to separate signal sources containing the mapping signal from those containing blood vessel artifacts, noise, etc. This is achieved by learning a linear demixing matrix. When applied to the observed image stack, the demixing matrix yields the estimated sources. Different methods exist for this learning; one of them, used in this work, is the Extended Spatial Decorrelation (ESD) approach. For this information about spatially shifted correlations of the mixtures is used.

There are different BSS techniques available, and [SSM+99] evaluated some of them on an image stack obtained during an ocular dominance experiment. It is obvious that, although able to extract activity maps, these algorithms have problems with sensor noise. One of the algorithms, which yielded the best results there, the ESD algorithm, is improved and applied to two data sets in this thesis.

The goal was to approximately decorrelate the estimated sources for several shifts instead of just one, as the ESD algorithm does. This decreases the influence of sensor noise on the separation results, and reduces the problem of selecting the right shift for decorrelation. For the optimization

of the extended error function an acccelerated gradient descent is used. Though this algorithm is dependent on the initialization of its parameters, it is more flexible in learning the demixing matrix, when compared to other multi-shift algorithms.

An artificial data set is used to control the sensor noise present in the analysed data. Issues analysed using the artificial data set are a comparison between single- and multi-shift algorithms, the differences in performance when using noise-robust sphering instead of the standard sphering approach (sphering is a preprocessing step needed by some algorithms, helpful to others), and the effects of spatially correlated sensor noise instead of white sensor noise. The results indicate a superior noise robustness of the algorithm developed in this thesis, when compared to other variants of the ESD algorithm. Evaluation for the second data set, the same ocular dominance experiment as in [SSM+99], shows, that the newly developed algorithm compares favorably to the other ESD variants and is very well able to extract ocular dominance maps. The extracted image maps have better separation of the mapping signal from other sources like blood vessel artifacts or global signal than other algorithms, which are currently used.

Chapter 3

Background

3.1 Optical Imaging of Intrinsic Signals

3.1.1 Overview of Optical Imaging

Optical imaging is a technique used to acquire information about the functional architecture of the brain. When it is used in imaging of the primary visual cortex, which is the case for the data analysed in this work, different visual stimuli are presented to the eye. These stimuli evoke neural activity in the primary visual cortex. Changes in light reflection of the cortical tissue, which are related to the activity of neurons, are used to extract activity maps of the cortex from camera recordings. These maps show which parts of the mapped regions are activated by the presentation of the stimulus. Examples are ocular dominance maps, indicating which regions of the visual cortex are excited by the left and which by the right eye, and iso-orientation maps, displaying regions excited by edges with a given orientation. Signals underlying the changes in light reflection, which are used for the creation of such maps, are changes in light scattering and amount of deoxygenated hemoglobin.

Bonhoeffer and Grinvald give a very good and concise introduction to optical imaging using intrinsic signals in their book chapter [BG96]. In the following I provide a summary to introduce the reader to this field.

The existence of intrinsic signals, which can provide information about the activity of neurons, is already known for many years. In 1949 Hill and Keynes reported about “Opacity changes in stimulated nerve” [HK49]. In 1986 these signals were reported to be used for the creation of cortical maps of neural activity in [GLF$^+$86]. Optical imaging using intrinsic signals currently provides very high spatial resolution, compared to other in vivo imaging techniques available, like fMRI. Though its temporal resolution is slower than the one achieved with voltage-sensitive dyes, the combination of still reasonable precision in time and a high level of spatial detail opens up many existing and new fields of application for this technique. An advantage when compared to voltage sensitive dyes is that the brain, and if less spatial resolution is sufficient also the dura, are not severed. No extrinsic substances are used which could damage the brain or change its function; only the skull and dura have to be opened.

Three kinds of maps are usually extracted in imaging sessions. Single condition images show which regions of the cortex are activated and which are inactive during presentation of one certain

stimulus (e.g. the reaction in the visual cortex to a moving grating presented to only one eye). Difference maps, on the other hand, basically take two single condition images for orthogonal stimuli[1] and use their difference or ratio to enhance the signal to noise ratio and eliminate regions which are active regardless of the stimulus. Several single condition images can be combined using various methods to obtain a map representing multiple stimuli; a color coded map with different colors for the different edge orientation preferences of cortical neurons is an example of this.

Maps created by optical imaging are used for many different applications: They are useful in the investigation of structural elements of the functional architecture of the brain; currently mainly the visual cortex of different animals is analyzed in this way, although other sensory systems are also investigated. Examples are experiments concerning the visual pathway ([TRS93]) and inferotemporal area (object recognition, [WTT94]). The feasibility of chronic experiments offers the possibility to analyze postnatal experience-dependent plasticity and the development of the neocortex over several weeks or months ([KB94, CB94, GB96]). Effects of environmental changes, e.g. monocular deprivation, can be explored. Using a special lens system (a macroscope) with very high numerical aperture for projection of the reflections onto the camera also makes it possible to focus the camera in different depths of the cortex; this works up to a depth of 0.5 to 1.0 mm into the surface of the cortex.

For neurosurgery optical imaging could one day offer the prospect to determine the borders of functional areas in the vicinity of surgical procedures, or the location of epileptic events. Imaging, though with much less spatial precision, is possible through the intact dura (cats, [FLTG90]) or even through the thinned skull (rats, [MKDF93]), when using infrared light. For monkeys recordings are usually done from the open brain. Research is currently done to make optical imaging through the intact skull of humans possible, as an aid in diagnosis and surgery preparation.

The following subsection explains some of the principles and signal types underlying optical imaging using intrinsic signals.

3.1.2 Signal Sources

The images recorded by optical imaging from the cortex of animals contain very small temporal and spatial changes in the level of light reflection. These changes are below the level a human can observe. The use of modern cameras (see description of experimental setup in the next section) with high signal to noise ratios allows to detect them. The detected change of light reflection is called the total signal in this thesis.

The total signal can be split into components with different statistical properties, e.g. time coureses. One of these is the mapping component, which is related to local neural activity and has a fine spatial resolution. Another part of the total signal, the global component, has a coarser spatial resolution and is not suitable for optical imaging. Further components include blood vessel patterns, vasomotor signal and ongoing activity. Each of the components has biophysical causes. In the following the components are explained in more detail, followed by a description of the underlying biophysical components.

The mapping component contains biophysical components, whose amplitude and spatial pattern correspond best to local neural activity. Its spatial resolution is precise enough to be used for

[1] Orthogonal stimuli are presumed to activate nearly disjunct populations of neurons. Right eye and left eye stimuli could be assumed to be orthogonal.

optical imaging. The underlying biophysical components are mainly the light scattering and the deoxyhemoglobin components.

The global component is comprised by signals with less spatial resolution. While its main underlying components, oxyhemoglobin concentration and blood flow and volume changes (e.g. local recruitment and dilation of venules), are still stimulus related (more oxygen is transported to regions with high neural activity), their time course and spatial resolution are coarser than that of the mapping signal.

Further components are interfering with the mapping component: The ongoing activity, spontaneous activation of the cortex ([ASGA96]), and the vasomotor signal, which is a slow oscillation of neural activity in the cortex ([MAZ$^+$96]) are examples of signals unrelated to the stimulus presentation. Larger vessels change their size and reflection of light due to changes in blood flow and volume, which causes artifacts which can be hard to separate from the mapping component.

The biophysical components (blood flow and volume changes, oxyhemoglobin, deoxyhemoglobin, and scattering components) are explained by Bonhoeffer and Grinvald in [BG96] as presented in the following paragraphs.

One biophysical component is the change in blood volume due to local capillary recruitment or dilation of venules. As a consequence the absorption of light by hemoglobin increases. This component is prevalent at 400 to 630 nm wavelength. At about 570 nm oxy- and deoxyhemoglobin have the same level of absorption and so, for this wavelength, the blood volume component dominates the total signal. A problem with this component is its spatial specificity. Blood flow changes normally affect rather large areas, meaning that other (deoxyhemoglobin and scattering) components are better suited for optical imaging.

Activity-dependent changes in oxygen saturation level of the hemoglobin are another component biophysical signal affecting optical imaging. The oxyhemoglobin and deoxyhemoglobin components are the changes in light reflection which are due to changes in the amount of oxygenated and deoxygenated blood in an area, respectively. Because oxyhemoglobin and deoxyhemoglobin have different absorption spectra, different time courses, and different spatial characteristics, the blood flow component has different effects on the total signal for varying wavelengths. The deoxyhemoglobin concentration increases during activation in a region because of the increase in oxygen consumption of active neurons. Contrary to this effect, the rush of fresh blood into active regions results in higher levels of oxyhemoglobin. These contradicting influences have different time series and whether the oxidation level increases or decreases depends on location and time. The deoxyhemoglobin signal constitutes about 30 to 50 % of the total signal; it starts about 200 ms after stimulus onset, rises during stimulus presentation and decays to baseline within 15 to 20 seconds after the end of the stimulus. The oxyhemoglobin signal is slower: It is constant or even reduced during the first 1.5 seconds of stimulus; then it rises for 1 to 3 seconds longer than the stimulus lasts. It only comprises about 5 % of the total signal. The deoxyhemoglobin signal is spatially the most precise (least smearing, about 100 μm) of all signal types mentioned and part of the mapping component, while the oxyhemoglobin signal is spatially and temporal less precise and belongs to the global component.

Another biophysical component arises from the light-scattering changes in regions of the cortex which are active. Ion and water movements, extracellular space dilations and contractions, swelling of subcellular compartments such as mitochondria, capillary expansion and neurotransmitter release all have effects on the scattering properties of neural tissue and are activity dependent. The scattering component becomes significant above 630 nm and is dominant in the near

infrared above 800 nm wavelength. The time course of this signal is reported to be very well suited for optical imaging: It is assumed to be wavelength independent, rises about 200 ms after stimulus onset and decays back to the baseline within 3 to 4 seconds after the stimulus stops. Its amplitude is about 10 % of the total signal. The smearing due to the scattering, which influences the spatial resolution achievable using this component, is estimated to be below 200 μm.

Bonhoeffer and Grinvald report that the properties of functional maps are generally very similar, regardless of the wavelength used for illuminating the brain during optical imaging. On the other hand, the mapping signal constitutes only about 5-10 % of the total signal at below 590 nm, whereas it is responsible for about 30-50 % at 605 nm. This is probably due to oximetry (level of oxidation of hemoglobin). The spatial resolution achievable by optical imaging using intrinsic signals is about 100 μm, limited by smearing and scattering effects of the neural tissue.

3.1.3 Experimental Setup

This section is intended to give a general overview of the experimental setup necessary for optical imaging. It is described in more detail in [BG96]. The actual experimental setup used for the optical imaging data, which is analyzed in this work, is described in chapter 6.

As a first step in optical imaging, generally a cranial window has to be mounted onto the skull of the animal. This involves trepanation of the skull of about 600 mm^2 (for cats) and fixation of a chamber on the skull, which is filled e.g. with silicon oil to protect the brain and provide good optical properties. Generally the dura is removed, too. But, especially for long term imaging it is advantageous to let it in place; good results have already been obtained this way, using infrared light (cats, [FLTG90]). Even totally non-invasive techniques or a thinning of the skull without a complete trepanation are used. It works well for rats ([MKDF93]); as a tool for diagnosis and surgery preparation for humans it is under being researched. In this case spatial resolution is lost and infrared light has to be used, because its absorption by the skull is much smaller for higher wavelengths.

Devices used for imaging are either CCD cameras or video cameras. Slow-scan CCD cameras provide a high signal to noise ratio (well capacity in relation to photon shot noise) and a high spatial resolution. On the other hand, their temporal resolution is poor, due to long exposure times for each frame. Modern video cameras also provide high spatial resolution. Differential imaging, where a reference frame (e.g. a blank exposure with no stimulus applied) can be subtracted (before digitization) from every frame, can be used to achieve better quality in digitizing images when using video cameras; all 8 bits are now available to encode the differences of pixels for the two images instead of the absolute value of each pixel. For both camera techniques, binning of neighboring pixels (spatially and temporal) can be used to improve the signal to noise ratio. For very low light levels cooled slow-scan CCD cameras are more appropriate, while in medium to high light environments video cameras achieve the better signal to noise ratio, because of their higher frame rates, which allow better temporal averaging.

A special arrangement of lenses, a macroscope, can be used to project the images onto the camera. It is essentially a microscope with very low magnification but very high numerical aperture and provides a very shallow depth of field. This allows to focus the camera into different depths of the cortex (a depth of up to 1 mm is possible). When focused below 300 μm the surface vasculature is blurred sufficiently for optical imaging.

During the recording of each optical image stack (except for blanks) one stimulus is presented. The stimulus set from which the stimulus is chosen depends on the maps which are to be created. For ocular dominance maps the stimuli could consist of gratings of all orientations moving in all directions; at these the animal looks with one and later with the other eye blocked. For orientation preference maps a single stimulus can be a number of moving edges with a certain orientation.

3.1.4 Extraction of Maps

Of the components mentioned in the last section only the mapping component, is suitable for optical imaging. This component has to be separated from all other components to obtain good maps of cortical activity. A problem is that other components (referred to as "biological noise" in the following), like the global signal, can also be related to neural activity and thus can show similar time courses. Furthermore, the images contain sensor noise, which is introduced by the camera; this noise is e.g. photon shot noise[2] or camera electric noise.

Photon shot noise and camera electric noise have high spatial frequency and are different from frame to frame. The higher the frame rate of the camera, the more photons per time unit are necessary to get statistically reliable photon counts (pixel values). The changes in reflectance due to activity, which are to be measured for optical imaging, are near 0.1 % of total reflectance. To achieve a signal to noise ratio (SNR) of e.g. 10, about 100,000,000 photons have to be counted in each frame.

Biological noise mostly has low spatial frequency (adjacent pixels are correlated) and temporal frequency (adjacent frames show similar noise). Signals like the vasomotor signal, the ongoing activity, and changes in blood-flow caused by breathing and heartbeat can have an amplitude which is much larger than that of the mapping component (depending on the animal, the experimental setup, whether the dura was removed, etc.). A synchronization of respiration of the animal and the start of recording with its heartbeat is very useful to minimize the noise cause by heartbeat and breathing: This way all experiments are done in the same phases of heartbeat and respiration. Physical movements also have to be taken into account. Shakes can move adjacent frames relative to each other and make a separation of the components even harder. Standard processing techniques for extraction of the activity maps from the recorded image frames, as presented in [BG96] are given in the following.

Single condition images are created by normalizing the images obtained under application of a single stimulus condition with a blank image. This blank image is either an image obtained when no stimulus was applied or a cocktail blank. The cocktail blank is a mixture of all images which were obtained for the complete set of stimuli. Latter has the advantage that regions, which are always active, and artifacts of blood vessels (which are larger in active regions), and the growth of active regions[3] are canceled out better. On the other hand, the creation of a cocktail blank needs some assumptions about what the complete set of stimuli may be. A complete set of stimuli is expected to uniformly activate the observed cortex region. The normalization can be a subtraction of the blank (whichever is used) or a division by the blank.

[2] The number of photons registered by the camera for a given light level is a stochastic process. I.e. for a given light level and recording time per frame the number of photons which is registered has a certain variance, introducing the noise. This is the larger the smaller the well capacity (number of photons per pixel the camera can accumulate before it overflows) of the camera is.

[3] Because of smearing effects active regions appear larger than they are in the image, by about 100 μm.

Difference maps are created by normalizing an image obtained for one stimulus with that of an orthogonal stimulus. The normalization can be either a subtraction of one from the other, followed by a division by a blank; or it can be a division of the image for one stimulus by that of the other. Both approaches give similar results, because the mapping signal, the changes in amplitude, is very small compared to the amplitudes (pixel values) themselves ([BG96]). The creation of difference images requires some assumptions, e.g. about the choice of orthogonal stimuli. A problem is that regions, which are active in both single condition images cannot be distinguished from regions, which are inactive in both.

A significant help in analyzing the recorded images can be first frame analysis. This procedure requires that, before presentation of the stimulus, one or more frames are recorded with no stimulus present. This frame (or their average, if more than one is used) is then subtracted from each of the following frames. Very slow biological noise can be canceled this way, although noise with high spatial frequency (photon shot noise) is increased, because the noise present in the first frame(s) is added to all other frames. This problem can be lessened if several frames without stimulus can be averaged for the subtraction.

Usually a bandpass filtering is used to extract the local mapping signal. The assumptions responsible for this choice of processing the images are based on the spatial power spectrum of the images. Components of high spatial frequency are assumed to be noise. This is realistic, because the spatial resolution of optical imaging techniques is limited to about 100 μm. Due to smearing and scattering effects features which are closer together cannot be distinguished. Thus anything finer than 100 μm must be noise. Components of low spatial frequency, on the other hand, are assumed to contain global signal components or other biological noise. Under these assumptions, a band of medium frequency used to filter the images obtained by optical imaging should give a good estimate of neural activity. But the highpass filtering of the images is questionable, as there is no fixed frequency separating the local and global signal components in the power spectrum. The statistics of pinwheel[4] distribution (density, number) in maps of orientation preferences can be changed if highpass filtering is applied to the images ([SOM$^+$97]). If bandpass filtering is applied to white noise it is possible to obtain images similar to orientation preference maps ([RS90]). In contrast to this, reasonable lowpass filtering does not change these statistics.

As a consequence of the shortcomings of bandpass filtering other techniques are evaluated for their use in optical imaging. The technique used in this thesis assumes a linear mixing model of the biophysical components: The mapping component, the global component, vasomotor signal, etc. are assumed to be added to the background image, weighted by their time course. Blind Source Separation algorithms, explained in the next section, are used to estimate a demixing matrix, which then allows to retrieve the components from the recorded mixtures.

3.2 Overview of Independent Component Analysis and Blind Source Separation

A problem often used to illustrate the Blind Source Separation (BSS) problem is the Cocktail Party problem. Imagine you are a guest at a cocktail party and there are several small groups of people all talking (in not too low voices) at the same time. Nevertheless most people are still able to understand what their conversation partner is saying. Translated to the BSS framework this is

[4] Pinwheels are locations with orientation singularities; all orientations are represented in the vicinity.

interpreted as following: Several sources (the voices of the people talking) are mixed (in the ear) to give two observed mixtures (the sounds "heard" in both ears).

The cocktail party problem illuminates several issues of the Blind Source Separation task. First the voices of people are convolved. This can occur for example due to echos reflected from the walls or by the acoustic reception in the ears. Convolution of the sources can possibly help in the separation process by introducing temporal dependence in the signals. On the other hand, to obtain the original sources, the convolution has to be inversed. This process has to be learned in addition to the unmixing.

Furthermore, due to different paths from the sources to the two sensors, the ears, propagation delays are occurring. Depending on the location of a speaker its voice can arrive earlier in one ear than in the other. This makes the mixing, and consequently the demixing, process more complicated. The demixing cannot be instantaneous, i.e. it cannot simply use the values of the mixtures at one point of time to recover the original sources for that point of time. Instead it has to remember past observations to take them into account during demixing.

A third point is that people only have two sensors (their ears) available which they can use for separation of a potentially unknown number of voices. For a linear mixture without assumptions about the underlying source signals at least the same number of sensors as sources is necessary to separate them all. On the other hand, the attention of people generally focuses on only one source, the mixtures need not be linear, and the sources are convoluted; all this could make the separation process easier.

The mixing process in the ear is potentially nonlinear. This poses the question of the appropriate demixing procedure which is the inverse to the mixing. In the brain the analysis of sounds heard by the ears seems to be adapted to the mixing process very well; if a demixing is to be done artificially some assumptions about the underlying mixture process are necessary.

Finally, the mixing process during a cocktail party is highly dynamic. People are walking around while talking, the listener moves his head, new people arrive and others become quiet. The demixing process has to adapt to all these changes, and the human brain and ears are obviously very good in doing that.

Current algorithms developed for the Blind Source Separation problem normally simplify many of these issues. First the mixing process usually is assumed to be linear. Furthermore, often time delays in the mixing process are ignored; this can be justified for real world data if the time delay of the arrival of the signals at the different sensors is shorter than the sampling rate or the temporal correlation of the signals is broad enough (i.e. the signals change slowly). In artificial data sets the mixing normally does not involve time delays, anyway.

Moreover, the number of sources is in general pretended to be at most as high as the number of sensors. Often both numbers are assumed to be the same. For a linear mixing process this allows inversion of the mixing, because for the number of sensors greater or equal to the number of sources the mixing matrix has full rank (in general, if it is not singular).

Another simplification often assumed is that the sources are not convolved before they are mixed. If a convolution has to be inversed information about the time series of the signals must again be taken into account, making the demixing more complicated.

Many algorithms were proposed for the task of blind source separation. In this thesis the Molgedey & Schuster algorithm proposed in [MS94], and evaluated for optical imaging in [SSM+99] (as the

ESD, Extended Spatial Correlation, method) is used and enhanced. It is based only on second order statistics (correlations and shifted correlations) of the data. Second order statistics is sufficient for BSS in the case that auto-correlation of the sources exist for non-zero shifts and that they are reasonably different between the sources. Further information and a mathematical formulation is given in section 4.

Other algorithms, which are often summarized under the label ICA (Independent Component Analysis) use higher order statistics of the data, statistical independence and probability density function (pdf.) assumptions of the sources, or non-linear PCA, among others, for blind source separation. Lee et al. [LGBS99] give a good overview of several Independent Component Analysis algorithms.

Two algorithms which minimize mutual information between estimated sources are described in [BS95] and [Gir97]. In the first paper Bell and Sejnowski give a learning rule for a neural network which maximizes information transfer and thus minimizes mutual information in the outputs, making them independent. In the second paper Girolami uses negentropy maximization (which is the Kulback-Leibler divergence between the pdf. of the source estimates and the Gaussian distribution with the same mean and variance) to minimize mutual information of the estimated sources.

A different approach is taken by Oja in [Oja97]. He analyses the convergence and source separation abilities of the nonlinear PCA algorithm. This is a modification of the network he developed for Principal Component Analysis ([Oja92]) to include nonlinear instead of linear neurons.

In [HO97] Hyvärinen et al. present a fast fixed point algorithm which optimizes the kurtosis (in statistics the diagonal of the fourth-order cumulant tensor of a probability density, see [Nik93]) in order to extract non-Gaussian independent components from the observed data. The kurtosis measures, how "peaked" and "long-tailed" a probability distribution is. Gaussians have kurtosis 0; the sharper the peak of a pdf and the longer its tails are, the higher its kurtosis is; flat distributions have negative kurtosis. If, after sphering the observations (transforming the mean to zero and the variance to one), a rotation is found which maximizes or minimizes the kurtosis in direction of all the axes, a complete separation is found.

The work in [SSM+99] indicates that the M & Schuster algorithm (referred to as Extended Spatial Decorrelation, ESD) is most appropriate for optical imaging data, when compared with the ICA algorithms given in [BS95, Ama96] and [HO97]. The spatial auto-correlation structure of this data is suited very well for blind source separation using the ESD algorithm. The data and its auto-correlation structure is presented in detail in section 6. Nothing special needs to be assumed about the probability density functions of the underlying sources (e.g. they do not need to have super- or sub-Gaussian distributions), which would be necessary to use other ICA algorithms.

A problem for all BSS algorithms is sensor noise, which cannot be modeled as a separate source, but is added after the mixing process. The experiments in [SSM+99] indicate that the M & S algorithm is the best of the evaluated algorithms in coping with noise. The development of a noise robust algorithm for BSS was one of the main goals of this thesis; the idea is that by using more information of the cross-correlation structure of the mixtures than the M & S algorithm does, it is possible to use the gained redundancy for canceling part of the sensor noise. The proposed algorithm is described in section 4.4.

Chapter 4

Algorithms

This chapter presents and explains some second order statistical approaches to the problem of Blind Source Separation. In section 4.1, after a short introduction to the used notation, the framework of Blind Source Separation, its assumptions and some comments concerning Blind Source Separation on optical imaging data are introduced. The first algorithm presented is the Extended Spatial Decorrelation algorithm published in [MS94, SSM$^+$99] (section 4.2). It is the basis for the two other algorithms evaluated in this thesis. The Jacobi Method is explained in section 4.3, followed by an accelerated gradient descent algorithm developed in this work (in section 4.4).

4.1 Blind Source Separation Problem

4.1.1 Mathematical Notation

In the following a short overview of often used notation is provided.

Vectors and matrices are printed in bold face, scalars in italics. A hat $\hat{\cdot}$ denotes an estimated quantity, e.g. estimated sources. Angle brackets $\langle\cdot\rangle_i$ express the average with respect to the given sample index i. $\mathbf{r}$ is a vector specifying a pixel in images, while $\Delta\mathbf{r}$ is the distance vector between two pixels (difference of their respective vectors). $\mathbf{s}(\mathbf{r})$ denotes a vector of sources at a location $\mathbf{r}$; similarly $\mathbf{y}(\mathbf{r})$ is a vector of mixtures, and $\mathbf{y}'(\mathbf{r})$ represent sphered mixtures. $\mathbf{C}^{(s)}(\Delta\mathbf{r})$ is a cross-correlation matrix of the sources for the given shift $\Delta\mathbf{r}$, and $\mathbf{C}(\Delta\mathbf{r})$ stands for the cross-correlation matrices of the mixtures. $\mathbf{A}$ and $\mathbf{W}$ denote the mixing and demixing matrices, respectively. Noise is expressed using the variable n.

The cross-correlation matrices of sources $\mathbf{C}^{(s)}(\Delta\mathbf{r})$ and of the mixtures $\mathbf{C}(\Delta\mathbf{r})$, for a certain shift $\Delta\mathbf{r}$ are defined as following:

$$\mathbf{C}^{(s)}(\Delta\mathbf{r}) = \Big\langle \mathbf{s}(\mathbf{r})\mathbf{s}^T(\mathbf{r}+\Delta\mathbf{r}) \Big\rangle_{\mathbf{r}} \tag{4.1}$$

$$\mathbf{C}(\Delta\mathbf{r}) = \Big\langle \mathbf{y}'(\mathbf{r})\mathbf{y}'^T(\mathbf{r}+\Delta\mathbf{r}) \Big\rangle_{\mathbf{r}} \tag{4.2}$$

The diagonal entries in these matrices are the auto-correlations. $C_{i,i}^{(s)}(\Delta\mathbf{r})$ is the auto-correlation of source i for shift $\Delta\mathbf{r}$. The off-diagonal elements are the cross-correlations between the sources or mixtures.

4.1.2 Blind Source Separation and Assumptions

The model

In the Blind Source Separation (BSS) framework the observed data is modeled as a set of observation vectors $\{\mathbf{y}\}$, which are a linear mixture of unobserved source vectors $\{\mathbf{s}\}$ using the mixing matrix $\mathbf{A}$. If sensor noise $\mathbf{n}$ is included in the model, it is added after the mixture. In this thesis spatial BSS is used, so the sample index for each source and mixture is the vector $\mathbf{r}$, denoting a pixel in a source prototype or mixture image:

$$\mathbf{y}(\mathbf{r}) = \mathbf{A}\mathbf{s}(\mathbf{r}) + \mathbf{n} \tag{4.3}$$

The goal of BSS algorithms is to find a demixing matrix $\mathbf{W}$, which gives source estimates $\hat{\mathbf{s}}(\mathbf{r})$, which are optimally decorrelated:

$$\hat{\mathbf{s}}(\mathbf{r}) = \mathbf{W}\mathbf{y}(\mathbf{r}) \tag{4.4}$$

Assumptions

First, the mixing process usually is modeled to be linear. Non-linear mixtures would need further assumptions about the underlying mixing model. Such extensions are hard to derive for optical imaging data (see [SO99]), which is the actual target application of the algorithm developed here. Furthermore, more parameters are normally necessary for non-linear models, making their estimation harder.

In this thesis BSS is performed using spatial shifts. The second assumption concerns the conditions the data must conform to in order for this method to work. One is that sources have non-zero auto-correlation functions $C_{i,i}^{(s)}(\Delta\mathbf{r})$, which differ among the sources for the shifts $\{\Delta\mathbf{r}\}$ used by the algorithm (R is the number of data points, over which the sample index $\mathbf{r}$ runs):

$$C_{i,i}^{(s)}(\Delta\mathbf{r}) = \langle s_i(\mathbf{r}) s_i(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}} = \frac{1}{R}\sum_{\mathbf{r}} s_i(\mathbf{r}) s_i(\mathbf{r}+\Delta\mathbf{r}) \tag{4.5}$$

At the same time, to make successful separation possible, the sources must have vanishing cross-correlation functions:

$$C_{i,j}^{(s)}(\Delta\mathbf{r}) = \langle s_i(\mathbf{r}) s_j(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}} = 0 \quad ;\forall\, j \neq i, \forall\, \Delta\mathbf{r} \tag{4.6}$$

The former condition, non-vanishing auto-correlations, mean that images are smooth: Neighboring pixels are not independently drawn from a probability density. Data, whose auto-correlation

vanishes for all non-zero shifts, is better processed with some ICA methods, which make other assumptions about the sources (see overview of ICA algorithms in section 3.2).

A third assumption is that the number of observed mixtures, i.e. the dimension of each $\mathbf{y}$, must be the same as the number of estimated sources, i.e. the dimension of each $\mathbf{s}$ (at least for the BSS algorithms used in this work). If the number of real sources is less, some estimated sources will contain mainly noise. If it is larger, the algorithms cannot separate all sources.

Issues concerning BSS on Optical Imaging Data

For BSS algorithms it is generally assumed that the sources are independent. Spatial second order BSS uses the fact that the sources should, for all spatial shifts, be uncorrelated, if they are statistically independent. In fact it is sufficient, if the sources are uncorrelated for those shifts which are used by the algorithm. The family of Primary Component Analysis (PCA) algorithms ([Oja92]) is only based on the assumption that cross-correlations for the zero-shift should vanish. That only constrains the space of solutions enough to recover sources rotated by an arbitrary angle. The correct sources can only be found, if the corresponding mixing matrix is symmetrical. BSS algorithms make further assumptions. One possibility is, as mentioned, that the sources are assumed to be uncorrelated with versions of other sources which are shifted by a certain amount in space (for spatial BSS).

The auto-correlation of sources must be non-zero, at least for the considered shifts, for this assumption to be useful (see figure 5.2 for an example of auto- and cross-correlations). The assumption of non-vanishing auto-correlations are very appropriate for the extraction of mapping signals from optical imaging recordings, because the neural activity which underlies them usually affects regions which span several pixels in the recorded images. Thus, neighboring pixels are correlated. The question of vanishing cross-correlations between the sources is less clear. The mapping signal should be uncorrelated with artifacts like blood vessels. On the other hand, it is not impossible that other sources, e.g. the global signal, could have similar correlation structures. Nevertheless, simulations performed in [SSM$^+$99] and during this work showed that the approach is very well suited for optical imaging.

In real world applications contamination of the data by noise has to be considered. This thesis will deal with two types of noise. The first type of noise are sources in which the experimenter is not interested. This type consequently is not treated specially, but modeled as sources, and BSS algorithms can separate them automatically. The second type of noise, here called sensor noise, is often not considered for BSS algorithms and a serious problem for many of them. In optical imaging this can be e.g. photon shot noise or camera electric noise. It is added after the mixing process and cannot be modeled as an own source. Although the Extended Spatial Decorrelation algorithm, as presented in [SSM$^+$99], performs better than other BSS approaches, it was an objective of this work to improve its noise robustness.

In the noiseless case (concerning sensor noise, $\mathbf{n} = \mathbf{0}$) the correct demixing matrix would be the inverse of the mixing matrix. In the noisy case, $\mathbf{W}$ has additionally to compensate for the added noise; the optimal demixing matrix decorrelates the estimated sources $\hat{\mathbf{s}}$, and is not necessarily the inverse of the mixing matrix any more. Furthermore, it is, even in the noiseless case, only possible to estimate a scaled and permuted version of the inverse using BSS algorithms.

For BSS, as it applies to optical imaging, two different approaches are possible, temporal and spatial BSS. In temporal BSS the elements of each source or observation vector is a time series,

for one source or one pixel, respectively. Correspondingly, in spatial BSS, which is used in the analyses in this thesis, every observation vector is an image, while the sources are spatial prototype patterns. $y_i(\mathbf{r})$ is the value of pixel $\mathbf{r}$ in the ith observed image frame. $s_i(\mathbf{r})$ respectively denotes the value of pixel $\mathbf{r}$ for source i. The observed mixture for time t results from the summation of the prototype patterns, weighted by the value of their respective time series at time t. The underlying assumptions for the choice of spatial or temporal BSS follow:

The smoothness of the sources is one criterion for the choice between spatial and temporal BSS. The more the auto-correlation structures of the sources differ, the better the separation performance one can expect. For image stacks the interpretation is: If the correlation structure of neighboring pixels is more prevalent than the correlation structure of the time series of the pixels, spatial BSS will be more promising.

Another factor of influence, especially for mixtures contaminated by high levels of sensor noise, is the number of samples for each mixture. Assuming a low number of frames with many pixels in each frame, as is the case for the data sets in sections 5 and 6, the number of observation vectors is higher if spatial BSS is chosen. The time series of each pixel is an observation vector, instead of the frames at different points in time (as for temporal BSS). If the data is very noisy, it is of much help to have many samples for each mixture, as it better allows to cancel out the noise by averaging.

A third topic to consider for the choice between spatial and temporal BSS is, especially for large data sets, the memory requirement of BSS algorithms. For the data sets used in this thesis, with many pixels in few frames, this also favors spatial BSS: The mixing and demixing matrices, as well as mixture and estimated source vectors, have smaller dimensionality. The number of samples, which is higher for spatial BSS, is less important, because an averaging takes place over all samples during calculation of the cross-correlation matrices (see algorithms in the following sections).

4.2 Spatial Molgedey & Schuster Algorithm

This algorithm is the basis of the other two algorithms explained later. It was published in [MS94]. There the authors used temporal correlations for the separation; in [SSM$^+$99] the algorithm was used to perform spatial Blind Source Separation and was applied to optical imaging data. The spatial version of the algorithm, which is used in this work, is called ESD (Extended Spatial Decorrelation). In the rest of this work I use this term to refer to the idea of using information about correlations for different shifts for separation.

4.2.1 Motivation and Characterization

This algorithm uses, besides the zero-shift cross-correlations $\mathbf{C}(\mathbf{0})$, one shifted cross-correlation matrix $\mathbf{C}(\Delta\mathbf{r})$ for computing the demixing matrix. Thus it is called a single-shift algorithm here, in contrast to the multi-shift algorithms, which are presented in the following sections. Because it does not not only use the zero shift for decorrelation, as PCA (Principle Component Analysis) does, it is called the *Extended* Spatial Decorrelation algorithm.

An advantage of this algorithm is that exact solutions for estimated sources and the demixing matrix can be obtained explicitly, by solving an Eigenproblem. There is no need to solve an

optimization problem iteratively. Thus it is very fast, after the correlation matrices have been calculated. Latter is necessary for all presented algorithms, so it is not taken into account for the comparison of algorithms.

4.2.2 Description

Similar to all algorithms presented in this thesis, the ESD algorithm optimizes a cost function, which diagonalizes correlation matrices of the estimated sources; in this case only for the zero- and one other shift $\Delta\mathbf{r}$:

$$\begin{aligned} E(\mathbf{W}) &= \sum_{i\neq j}\left(\left(\mathbf{W}\mathbf{C}(\mathbf{0})\mathbf{W}^T\right)_{i,j}\right)^2 + \left(\left(\mathbf{W}\mathbf{C}(\Delta\mathbf{r})\mathbf{W}^T\right)_{i,j}\right)^2 \\ &= \sum_{i\neq j}\langle\hat{s}_i(\mathbf{r})\hat{s}_j(\mathbf{r})\rangle_{\mathbf{r}}^2 + \langle\hat{s}_i(\mathbf{r})\hat{s}_j(\mathbf{r}+\Delta\mathbf{r})\rangle_{\mathbf{r}}^2 \end{aligned} \tag{4.7}$$

For the calculation of the cross-correlation matrices the data is first sphered. The standard sphering procedure is

$$\mathbf{y}'(\mathbf{r}) = \mathbf{D}\mathbf{y}(\mathbf{r}) \quad \text{, where} \tag{4.8}$$

$$\mathbf{D} = \left\langle \mathbf{y}(\mathbf{r})\mathbf{y}^T(\mathbf{r})\right\rangle_{\mathbf{r}}^{-1/2} \tag{4.9}$$

Here $\mathbf{y}'$ is the sphered data; $\mathbf{D}$ is the sphering matrix, which transforms the data to have variance 1 along all axes. Before applying this step the data must be shifted to have zero mean. Then the cross-correlation matrices for the two shifts have to be computed:

$$C_{i,j}(\mathbf{0}) = \left\langle y_i'(\mathbf{r})y_j'(\mathbf{r})\right\rangle_{\mathbf{r}} = \mathbf{I} \tag{4.10}$$

$$C_{i,j}(\Delta\mathbf{r}) = \left\langle y_i'(\mathbf{r})y_j'(\mathbf{r}+\Delta\mathbf{r})\right\rangle_{\mathbf{r}} \tag{4.11}$$

The former cross-correlation matrix is the identity matrix here, because of the previously applied sphering. In [MS94] Molgedey and Schuster only require the sources to have zero mean; other algorithms need the data to be sphered, or are more stable for such preprocessing, so I generally apply it to the data.

Now the Eigenvalue problem

$$\mathbf{C}(\mathbf{0})\mathbf{C}^{-1}(\Delta\mathbf{r})\hat{\mathbf{A}} = \hat{\mathbf{A}}\mathbf{\Lambda}(\mathbf{0})\mathbf{\Lambda}^{-1}(\Delta\mathbf{r}) \tag{4.12}$$

can be solved for the estimated mixing matrix $\hat{\mathbf{A}}$. This step is further commented in appendix A.1. $\mathbf{\Lambda}(\mathbf{0})$ and $\mathbf{\Lambda}(\Delta\mathbf{r})$ are diagonal matrices with the Eigenvalues. $\hat{\mathbf{A}}^{-1}$ then diagonalizes as well $\mathbf{C}(\mathbf{0})$ as $\mathbf{C}(\Delta\mathbf{r})$ and can be used to recover the unknown sources:

$$\hat{\mathbf{s}}(\mathbf{r}) = \hat{\mathbf{A}}^{-1}\mathbf{y}'(\mathbf{r}) \tag{4.13}$$

One drawback of this algorithm is the question of how to choose the single shift which is used by this algorithm. Molgedey and Schuster do not give a guide for this choice. In simulations presented later we used a variant of the M & S algorithm, which specifies this choice; the definition is given in section 4.2.3.

4.2.3 Variations

Heuristical choice of shift

On open question for the ESD algorithm is how to choose the arbitrary shift. Molgedey and Schuster only require it to be chosen such that

$$C_{i,i}(\mathbf{0})C_{j,j}(\Delta\mathbf{r}) \neq C_{i,i}(\Delta\mathbf{r})C_{j,j}(\mathbf{0}) \qquad ;\forall i \neq j \tag{4.14}$$

My experience is that, at least in presence of sensor noise, the choice of shift can have a crucial influence on the separation performance.

The heuristic presented here has the goal to maximize the signal to noise ratio for the components of the cross-correlation matrix used. From a set of possible shifts the one with the largest off-diagonal entries in the corresponding cross-correlation matrix is used, in the hope that the noise level is the same in all correlation matrices:

$$\Delta\mathbf{r}_{cor} = \text{argmax}_{\{\Delta\mathbf{r}\}} \frac{\text{norm}\left(\mathbf{C}(\Delta\mathbf{r}) - \text{diag}\left(\mathbf{C}(\Delta\mathbf{r})\right)\right)}{\text{norm}\left(\text{diag}\left(\mathbf{C}(\Delta\mathbf{r})\right)\right)} \tag{4.15}$$

The norm used here calculates the largest singular value of its argument (the MATLAB function `norm`), and diag$(\cdot)$ sets all off-diagonal elements of its argument matrix to zero. The numerator calculates the norm of the off-diagonal elements of the correlation matrix, which is then normalized by the norm of the diagonal elements. Now the question is how to choose the set of shifts $\{\Delta\mathbf{r}\}$ which are examined. In general, not all possible shifts can be examined; for large data sets, the computation of all cross-correlation matrices would take (too) much time. An observation made by us is that the value of the heuristic is more or less smooth over the shifts, for the data sets evaluated in this thesis. So it seems reasonable to spread out some shifts among the set of possible ones. Sections 5 and 6 present a possible choice and results for them.

This version of BSS is called `cor` in the simulation results sections of this thesis.

Optimal Shift

For the artificial data in section 5 a comparison of the estimated sources with the real sources is possible, because the latter are known. The best separation result possible using only a single shift can be determined in this case by performing a BSS for each possible shift. The function used to

determine the quality of the separation, i.e. the function comparing real and estimated sources, is the following:

$$\begin{aligned} \mathsf{RE}(\mathbf{W}) &= \text{offdiag}\left(\sum_{\mathbf{r}} \hat{\mathbf{s}}(\mathbf{r})\mathbf{s}^T(\mathbf{r})\right) \quad \text{, with (see F1 in [KO99])} \qquad (4.16) \\ \text{offdiag}(\mathbf{C}) &= \frac{1}{N}\sum_i \frac{1}{N-1}\left(\sum_j \frac{|C_{i,j}|}{\max_k |C_{i,k}|} - 1\right) \end{aligned}$$

In computation of the reconstruction Error (RE) first the correlations between estimated and real sources are calculated. In case of a successful separation (if the separation was unsuccessful then Inf is returned as reconstruction error) the resulting matrix should be close to a permutation matrix. Then the size of the non-permutation elements of this matrix compared to the permutation elements is computed. The smaller this ratio is, the better the separation performance. For a perfect demixing the RE is 0, if the original sources are uncorrelated. Otherwise the RE can be lower for the estimated sources than for the original ones (see separation example in results section, figure 5.3). A separation is counted to be unsuccessful, if the correlation matrix of estimated and real sources is not approximately a permutation matrix; this is the case if, after normalizing the rows of the matrix so that their largest element is 1, any column's number of "1"s is zero.

Simulation runs computing the optimal single shift and returning its results (demixing matrix, estimated sources) are denoted by opt in the toy data section 5.

Average Shift

The previous two algorithms (cor and opt) allow a comparison of the heuristical shift with the optimal single one. The mean algorithm is intended to provide an estimate of the quality a randomly selected shift could be expected to give. For each shift, as opt does, the reconstruction error is computed. Instead of taking the minimum of all successful results they are summed up and divided by the number of successful shifts, giving an average reconstruction error of successful separation runs.

4.3 Jacobi Method

4.3.1 Motivation and Characterization

In the literature a procedure is known for the approximate simultaneous diagonalization of several matrices. [BGBM93] presents the idea. Several elementary rotations, Jacobi rotations, are used to build a rotation matrix which approximately diagonalizes the system. [CS96] gives explicit formulas for calculating the elementary rotations, and [ZM98] gives results for an application of this idea to Blind Source Separation. The Jacobi method assumes that only a rotation matrix is necessary for the approximate diagonalization. The advantage of this assumption is that a fast method for computation of the elementary rotations is available. The assumption is realistic in the

noiseless case. Then the data can be perfectly sphered, and only a rotation needs to be done to align the independent components with the axes.

The possibility to simultaneously diagonalize several matrices allows the use of multiple shifts for BSS problems. This has the advantage that there is no dependence on the quality of a single shift. By the approximate diagonalization the influence of noise is minimized. Furthermore, the selection of a single shift is no longer a problem, instead a collection of several shifts can be used, which are less critical to select. A discussion of the shifts selected follows in the chapters about the data sets.

4.3.2 Description

The sphering (calculation of $\mathbf{y}'$) and computation of the cross-correlation matrices $\mathbf{C}(\Delta\mathbf{r})$ has to be done as for the cor algorithm in section 4.2. But for this algorithm the cross-correlation matrices have to be computed for a set of shifts $\{\Delta\mathbf{r}\}$, not only for two shifts:

$$\begin{aligned} \mathbf{D} &= \left\langle \mathbf{y}(\mathbf{r})\mathbf{y}^T(\mathbf{r}) \right\rangle_{\mathbf{r}}^{-1/2} \\ \mathbf{y}'(\mathbf{r}) &= \mathbf{D}\mathbf{y}(\mathbf{r}) \\ C_{i,j}(\Delta\mathbf{r}) &= \left\langle y_i'(\mathbf{r}) y_j'(\mathbf{r}+\Delta\mathbf{r}) \right\rangle_{\mathbf{r}} \end{aligned} \tag{4.17}$$

The data has to be shifted to have zero mean before applying these steps.

In my experience the choice of the set of shifts is not very crucial; it is not always possible to include all possible shifts in this set, because of computation time. To include only the zero and one other shift, on the other hand, makes the choice very sensitive, as the results for the M & S algorithm in section 5 show. For a pattern in which several shifts are spread out among the possible shifts, generally good results were obtained. The exact choice for the set of shifts used in this thesis is explained for the data sets in the next two chapters.

The cost function minimized by the Jacobi algorithm is the same as for the dpa algorithm presented in the next section; but here the demixing matrix $\mathbf{W}$ is restricted to be orthogonal:

$$\begin{aligned} E(\mathbf{W}) &= \sum_{\Delta\mathbf{r}} \sum_{i \neq j} \left(\left(\mathbf{W}\mathbf{C}(\Delta\mathbf{r})\mathbf{W}^T \right)_{i,j} \right)^2 \\ &= \sum_{\Delta\mathbf{r}} \sum_{i \neq j} \langle \hat{s}_i(\mathbf{r}) \hat{s}_j(\mathbf{r}+\Delta\mathbf{r}) \rangle_{\mathbf{r}}^2 , \end{aligned}$$

For the approximate simultaneous diagonalization of the $\mathbf{C}(\Delta\mathbf{r})$ elementary rotation matrices $\mathbf{R}(i,j,c,s)$ are computed for all $i \neq j$, to optimize the cost function along all rotation axes. The matrices $\mathbf{R}(i,j,c,s)$ are the are equal to the identity matrix $\mathbf{I}$, except for the entries

$$\begin{pmatrix} R_{i,i} & R_{i,j} \\ R_{j,i} & R_{j,j} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \qquad ; \text{with } c^2 + s^2 = 1 \tag{4.18}$$

$\mathbf{R}(i,j,c,s)$ represents a rotation around the i-j plane, where c and s are the sine and cosine, respectively, of the rotation angle. c and s are computed to minimize the cost function (see [CS96]):

$$
\begin{aligned}
c &= \sqrt{\frac{x+r}{2r}} && (4.19)\\
s &= \frac{y}{\sqrt{2r(x+r)}} && (4.20)\\
r &= \sqrt{x^2+y^2}
\end{aligned}
$$

Here $[x,y]^T$ is any eigenvector associated to the largest eigenvalue of

$$
\begin{aligned}
\mathbf{G} &= \sum_{\{\Delta\mathbf{r}\}} \mathbf{h}(\mathbf{C}(\Delta\mathbf{r}))\mathbf{h}^T(\mathbf{C}(\Delta\mathbf{r})) \qquad \text{, with} && (4.21)\\
\mathbf{h}(\mathbf{C}) &= [\mathbf{C}_{i,i}-\mathbf{C}_{j,j}, \mathbf{C}_{i,j}+\mathbf{C}_{j,i}]^T
\end{aligned}
$$

After one elementary rotation is calculated, it is applied to the data set and the next elementary rotation is computed. After one iteration through all elementary rotations for $i \neq j$ the diagonalization generally is still improvable; thus all elementary rotations are computed again, until the change in the value of the cost function is below a threshold. It usually takes 1 to 2 iterations through all elementary rotations to be close to the optimum and about 3 to 10 iterations to converge for the data sets used in this thesis.

The algorithm presented in this section is denoted by jac0 in the following.

4.3.3 Alternative Sphering

The Jacobi algorithm is very sensitive to the sphering preprocessing, as it can only find rotational demixing matrices after the sphering is done. For data which is contaminated with sensor noise this requires careful analysis of the noise and then the choice of an appropriate sphering method.

Besides the standard sphering method presented for the cor and jac0 algorithms, [MPZ99] describe a modified sphering technique, which improves this algorithm greatly: The use of a shifted cross-correlation matrix for sphering is proposed. This approach should cancel out noise, as long as the spatial correlation of the noise is less than the shift $\Delta\mathbf{r}$ used in calculating the cross-correlation matrix. In the interest of a good separation result the shift should, on the other hand, be in the range where the real sources still show reasonable auto-correlation.

Instead of the sphering matrix $\mathbf{D}$ in equation 4.17 this sphering method uses an estimate of the real sphering matrix $\mathbf{D}_0$ computed by:

$$
\begin{aligned}
\hat{\mathbf{D}}_0 &= \left\langle \mathbf{y}(\mathbf{r})\mathbf{y}^T(\mathbf{r}+\Delta\mathbf{r}) \right\rangle_{\mathbf{r}}^{-1/2} && (4.22)\\
&= \left(\mathbf{A} \left\langle \mathbf{s}(\mathbf{r})\mathbf{s}^T(\mathbf{r}+\Delta\mathbf{r}) \right\rangle_{\mathbf{r}} \mathbf{A}^T + \left\langle \mathbf{n}(\mathbf{r})\mathbf{n}^T(\mathbf{r}+\Delta\mathbf{r}) \right\rangle_{\mathbf{r}} \right)^{-1/2}
\end{aligned}
$$

$$\begin{aligned} &\approx \left(\mathbf{A}\left\langle \mathbf{s}(\mathbf{r})\mathbf{s}^T(\mathbf{r}+\Delta\mathbf{r})\right\rangle_{\mathbf{r}}\mathbf{A}^T\right)^{-1/2} \\ &\approx \left(\mathbf{A}\left\langle \mathbf{s}(\mathbf{r})\mathbf{s}^T(\mathbf{r})\right\rangle_{\mathbf{r}}\mathbf{A}^T\right)^{-1/2} \end{aligned}$$

The first approximation assumes that noise has no auto-correlation and thus the expectation of the scalar product is zero. The second approximation uses the smoothness of the sources; if the auto-correlation of the sources is very strong, at least for very small shifts, this approximation is reasonable. $\hat{\mathbf{D}}_0$ is not necessarily positive definite any more. But nevertheless the success of this approach is shown during the evaluation of the algorithms for the artificial data set in section 5.2.

The dpa0 algorithm combined with this noise-robust sphering method is denoted by dpa*, where the star is the length (number of pixels) of the shift which is used for sphering. This is relevant in section 5.2.4, where different sphering shifts are evaluated.

Depending on the size of the shift used in calculation, this variant of the algorithm is called jac*, where the star is the length of the shift (in pixels).

4.4 Optimization by Gradient Descent

4.4.1 Motivation and Characterization

In the beginning the goal for this thesis was to develop a more noise robust version of the ESD algorithm, because analysis and experience with the cor, and also with the jac0 algorithm, indicated that these had problems in dealing with sensor noise. Three ideas are used to achieve this goal. The advantage this algorithm has compared to the M & S algorithm is the use of cross-correlation matrices for several shifts, in contrast to only two used by the M & S algorithm. This provides more information about the auto-correlation functions of the mixtures, because they are evaluated at many, instead of two, shift vectors. The demixing matrix $\mathbf{W}$ is over-determined by the diagonalization equations

$$\mathbf{W}\mathbf{C}(\Delta\mathbf{r})\mathbf{W}^T = \mathbf{\Lambda}(\Delta\mathbf{r}) \qquad ,\forall\Delta\mathbf{r} \tag{4.23}$$

(the $\mathbf{\Lambda}(\Delta\mathbf{r})$ are diagonal matrices). An approximate simultaneous diagonalization makes it possible to cancel out noise through the use of redundancy in the over-determined system. Furthermore, the algorithm does not rely on the quality of a single shift (the separation ability of the M & S algorithm depends critically on it). Instead it computes a solution which decorrelates the sources for all used shifts as good as possible. Section 5.2 shows that multiple shift algorithms can be better than the best single shift.

The second idea is to use a modified sphering technique (section 4.3.3), which is relatively robust against noise. It is published in [MPZ99]. This should not be important for the basic M & S algorithm, because it finds the optimum for the cost function regardless of whether the mixtures are whitened or not. But both the Jacobi algorithm and the gradient descent method gain very much from the use of the modified sphering technique. The Jacobi method can only find rotation matrices and thus depends on a good sphering, while the optimization process used here for simultaneous diagonalization by the accelerated gradient descent becomes more stable using this technique.

Third, the gradient descent algorithm does not restrict the demixing matrix in the same way the Jacobi method does, i.e. to be orthogonal. A non-orthogonal $\mathbf{W}$ allows better separation even in cases when the sphering does not work perfectly. This can be important in cases where the statistics of the noise are unknown and an appropriate sphering technique cannot be chosen.

The use of multiple shifts for the calculation of a suitable demixing matrix has another advantage, as mentioned in the section about the Jacobi algorithm: It lessens the problem of the choice of shift(s).

4.4.2 Description

The first step for this algorithm is again sphering. In theory it should not be necessary to sphere the data, as this algorithm is not restricted to find only rotation matrices as solutions for the demixing matrix (unlike the Jacobi algorithm). But experience shows that stability of convergence improves much if sphering is done as a preprocessing step; otherwise this algorithm often does not find a good separation, at least for the optical imaging data. The standard sphering procedure is given in section 4.3, it remains the same for this algorithm.

Next, the cross-correlations of the (sphered) mixtures $\mathbf{y}'$ for a preselected set of shifts $\{\Delta\mathbf{r}\}$ have to be computed (the same remarks as for the jac algorithms apply concerning the choice of shifts $\{\Delta\mathbf{r}\}$):

$$C_{i,j}(\Delta\mathbf{r}) = \left\langle y_i'(\mathbf{r}) y_j'(\mathbf{r}+\Delta\mathbf{r}) \right\rangle_{\mathbf{r}} \tag{4.24}$$

Now an iterative minimization of the cost function, given already in section 4.3 is done:

$$E(\mathbf{W}) \quad = \quad \sum_{\Delta\mathbf{r}} \sum_{i \neq j} \left(\left(\mathbf{W}\mathbf{C}(\Delta\mathbf{r})\mathbf{W}^T \right)_{i,j} \right)^2 \tag{4.25}$$

This minimization (for parameters $\mathbf{W}$) is performed by gradient descent. This adapts $\mathbf{W}$ to minimize the cross-correlations of the estimated sources, by approximate simultaneously diagonalizing all selected cross-correlation matrices. The derivatives for the cost function, used for the gradient descent in the beginning, are given in appendix A.2. In practice it proved later to be faster and sufficiently accurate to compute the derivative numerically, by the forward difference formula. This was done for all simulations presented in later sections.

The minimization must have a constraint to prevent the demixing matrix from converging to the zero matrix.[1] This gradient descent procedure uses the restriction of $\mathbf{W}$ to

$$(\mathbf{W}^{-1})_{i,i} = 1 \qquad , i = 1, \ldots, N \tag{4.26}$$

following the example in [MS94]. Molgedey and Schuster compare their algorithm with a recurrent neural network implementation. This network consists of a single layer of linear neurons,

[1] That would be a minimum of the cost function, because then all off-diagonal elements in the cross-correlation matrices for the estimated sources would be zero.

which have inhibitory connections among themselves. $\mathbf{T}$ is the matrix of these connections. The neurons have no self-feedback (or -inhibition), and so the matrix $\mathbf{T}$ has zeros in its main diagonal. If the signals have a slow rate of change (compared to the network dynamics), the architecture can be transformed into a feedforward network, whose input weights are given by

$$\mathbf{W} = (\mathbf{I} + \mathbf{T})^{-1}. \tag{4.27}$$

The constraint in equation 4.26 results from the fact that the neurons have no self feedback, i.e. the inverse of $\mathbf{W}$ always has ones in its main diagonal. Although this connection to recurrent neural networks is unimportant for this work, the constraint is reasonable and gives good results.

4.4.3 Acceleration by Conjugate Gradient Method

To improve and speed up convergence of the gradient descent it is combined with an acceleration technique. The conjugate gradient method described in *"Numerical Recipes in C"* [PFTV88] provided good results. In iteration t it calculates the Polak-Ribiere Conjugate Gradient direction $\mathbf{d}^t$ and uses its normalized version $\mathbf{g}^t$ as minimization direction:

$$\mathbf{g}^{t+1} = \mathbf{d}^{t+1}/|\mathbf{d}^{t+1}| \tag{4.28}$$

$$\mathbf{d}^{t+1} = \nabla E(\mathbf{W}^t) + \beta^{t+1}\mathbf{d}^t \quad \text{, where} \tag{4.29}$$

$$\beta^{t+1} = \frac{(\nabla E(\mathbf{W}^t) - \nabla E(\mathbf{W}^{t-1}))\, \nabla E(\mathbf{W}^t)}{(\nabla E(\mathbf{W}^{t-1}))^2}$$

$\nabla E(\mathbf{W})$ is the gradient of the cost function at location $\mathbf{W}$. In iteration t the minimum of the cost function is searched in direction $\mathbf{g}^t$. Instead of using the line-search algorithm described in [PFTV88] we use a dynamic stepwidth adaptation algorithm described in the next section to approximately find the minimum in the given direction.

The initialization for the parameters is $\beta^1 = 0$ and $\mathbf{d}^0 = \mathbf{0}$.

4.4.4 Dynamic Stepwidth Adaptation

The combination of the Polak-Ribiere rule with Stable Dynamic Parameter Adaptation was published in [Rüg96]. It uses information about the cost function at a few points to estimate a good stepwidth η^t for the current descent direction in iteration t. This is done by either multiplying or dividing the previous stepwidth η^{t-1} by a certain factor ζ, depending on which yields a smaller cost function value. ζ is a constant > 1 which can be arbitrarily chosen. The simulations presented later used $\zeta = 2.0$. If the cost for the current parameter set $\mathbf{W}^t$ is less than the one for the larger stepwidth, a special rule is applied. This decreases the value of the cost function in places where the cost function surface is nearly quadratic, i.e. close to a minimum.

Using the definition

$$e(\eta) = E(\mathbf{W}^t - \eta \mathbf{g}^{t+1}) \tag{4.30}$$

we can give the rule determining the used stepwidths:

$$\eta^{t+1} = \begin{cases} \eta^* & ;\text{if } e(0) < e(\eta^t\zeta) \\ \eta^t/\zeta & ;\text{if } e(\eta^t/\zeta) \le e(\eta^t\zeta) \le e(0) \\ \eta^t\zeta & ;\text{otherwise} \end{cases} \text{, where} \tag{4.31}$$

$$\eta^* = \frac{\eta^t\zeta/2}{1 + \frac{e(\eta^t\zeta)-e(0)}{\eta^t\zeta\mathbf{g}^{t+1}\nabla E(\mathbf{W}^t)}}$$

Using this stepwidth the update of the parameter set is:

$$\mathbf{W}^{t+1} = \mathbf{W}^t - \eta^{t+1}\mathbf{g}^{t+1} \tag{4.32}$$

The algorithm which combines minimization of the cost function by conjugate gradient descent with dynamic parameter adaptation is denoted by dpa0 in this thesis.

4.4.5 Alternative Sphering

Sphering can be very unreliable if sensor noise is present in the observations. To improve stability of dpa0 by more noise robust sphering, the same approach as in section 4.3.3 is used. Instead of the correlation matrix for the zero-shift it uses a shifted cross-correlation matrix for sphering, because these are less noisy. For data with broad auto-correlation functions this method should give a good estimation of the correlation matrix of the real unnoisy mixtures. The dpa0 algorithm combined with this noise-robust sphering method is denoted by dpa*, where the star stands for the length (number of pixels) of the shift-vector which is used for sphering. I.e. with a sphering shift of $\Delta\mathbf{r} = (1, 0)$ this would be dpa1. This is relevant in section 5.2.4, where different sphering shifts are evaluated for spatially correlated noise.

4.5 Simulation Times

All simulations were done on Sun Ultra5/10/30 Workstations. Table 4.1 gives approximate times needed for different steps of the algorithms, for both datasets examined in this work. The toy data set has three mixtures, the optical imaging data set has seven; both contain images with 256 $\times$ 256 pixels. The calculation of the cross-correlation matrices for the shifts in the star-pattern-set is necessary for all algorithms (cor, jac*, and dpa*). For the computation of the optimal (opt) and mean shift (mean) all cross-correlations in the examined 61×61 square were calculated, which took much more time. Without the time needed to compute cross-correlations, the cor algorithm is by far the fastest, followed by jac*. dpa* consumes most processor time, because of the iterative optimization. But, except for the calculation of the optimal shift (which is not applicable in practice, when the sources are unknown, anyway), all algorithms are suitable for interactive work. Further comments on the computational cost of the Jacobi algorithm can be found in [ZM98].

Algorithm	**Toy Data Set**	**OI Data Set**
Cross-correlations for opt algorithm (3721 shifts)	850 sec.	—
Cross-correlations for other algorithms (49 shifts)	13 sec.	47 sec.
cor	0.02 sec.	0.04 sec.
jac	0.5 sec.	3.6 sec.
dpa	12 sec.	15-49 sec.

Table 4.1: Approximate simulation times on a SunU10 needed by different algorithms for the artificial and optical imaging (OI) data sets presented in chapters 5 and 6. The cross-correlations for 3721 shifts are not computed for the OI data set, because the optimal shift cannot be determined (the original sources are unknown). The time for the dpa algorithm is variable, because for the difference stack it often reaches its termination condition (the value of the cost function, equation 4.25, is less than 0.005) before the number of maximal iterations allowed before stopping the algorithm. Latter is set to 100.

Chapter 5

Results on Toy Data Set

This thesis emerged from other projects concerned with processing of optical images. Among others the ESD algorithm was evaluated in these projects. It became clear that sensor noise in the data could be a serious problem. Consequently, one goal of this thesis was to find a way to make the ESD algorithm more robust against sensor noise. To make analysis of noise robustness possible, an environment where sensor noise could be controlled was created; the artificial data set used for the analysis is presented in section 5.1. Results for the different algorithms explained in chapter 4 are given in section 5.2. The performance on the original optical imaging data is shown in chapter 6.

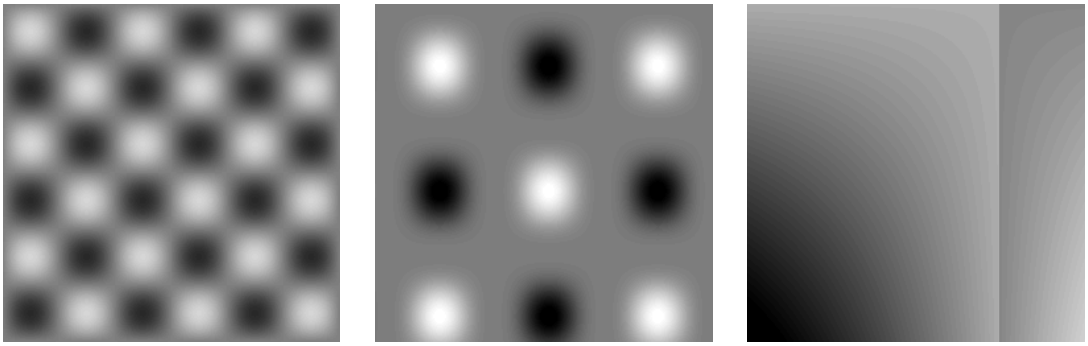
5.1 Data Set

The Molgedey-Schuster (ESD) algorithm for BSS and its variants require the sources to be uncorrelated; shifted as well as unshifted cross-correlations have to be (close to) zero for a successful separation of the mixtures. Furthermore, the sources have to be smooth, i.e. they have to have non-vanishing auto-correlation, at least for some shifts. A data set designed to conform to these requirements, containing three sources, is shown in figure 5.1. It is the same data set which was used in [SSM+99] for the evaluation of noise robustness. The sources are two two-dimensional sine-patterns, which are kind of similar to the patchy structure of orientation preference maps; the third source is intended to imitate biological noise, e.g. the gradient of oxygenation. All sources are normalized to a variance of 1. Below, in figure 5.2, a slice through the cross-correlation functions for the sources, for different horizontal shifts, is shown. The auto-correlation functions are smooth (graphs in the diagonal), while the cross-correlations are nearly vanishing.

Mixtures are created from the sources by applying a randomly generated 3×3 mixing matrix, using Gaussian random numbers with variance 1. The pixel sequence[1] of the sources is multiplied by the mixing matrix to yield the pixel sequence of the mixtures. This is done for every pixel. The mixing matrices used in this thesis usually had condition numbers between 3 and 10.

Generally, white Gaussian noise with a given noise level (standard deviation of noise) was added to the mixtures. For the simulations with correlated noise, white noise is spatially blurred using a Gaussian of radius 1, scaled to the given noise level, and subsequently added to the mixtures. The

[1] The sequence of the values of the same pixel in all three images.



Source 1 Source 2 Source 3

Figure 5.1: The set of three patterns for approximately uncorrelated sources. Mixtures of these were used in analyses of noise robustness of BSS algorithms.

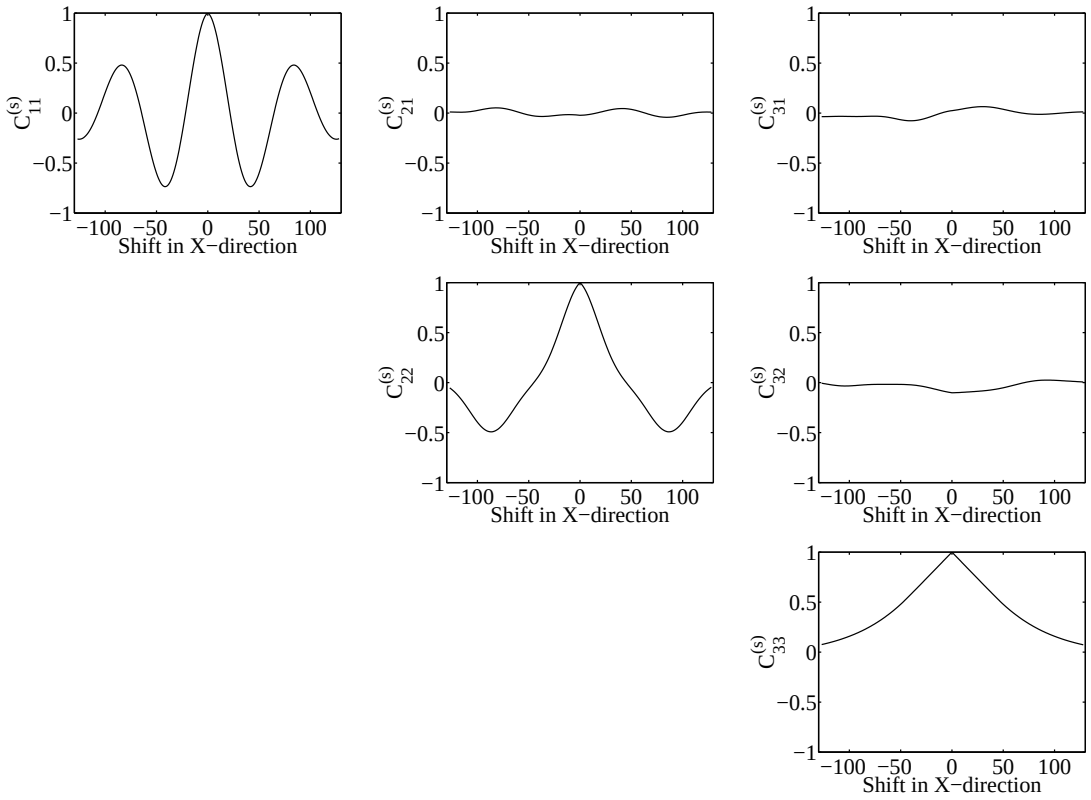


Figure 5.2: Correlations of the sources in figure 5.1. Auto-correlations are shown in the diagonal (from top left to bottom right for sources 1, 2, and 3), while their cross-correlations are given above that diagonal. Only shifts along the X-axis of the images are shown.

noise level (standard deviation) is converted to the Signal to Noise Ratio (SNR), measured in dB, for the plots. First the standard deviation of each mixture is calculated; their maximum σ_{data} is used in following formula:

$$\text{SNR} = 10 \log_{10} \frac{\sigma^2_{\text{data}}}{\sigma^2_{\text{noise}}} \tag{5.1}$$

For noise levels with standard deviations between 0 and 3.0 this resulted in SNRs between 30 and -5 dB, also depending on the mixing matrix. The higher the dB value of the SNR, the less noise is present in the data. For a SNR of 0 the largest signal (mixture without noise) and the noise have about the same amplitude (i.e. variance).

The separation performance is measured by the Reconstruction Error (RE), given in equation 4.16. If $\sum_{\mathbf{r}} \hat{\mathbf{s}}(\mathbf{r})\mathbf{s}^T(\mathbf{r})$ is not approximately a permutation matrix,[2] the separation is counted as a failure (Inf (Infinity) is returned). Otherwise it is a success and the RE is the normalized sum of the absolute values of the non-permutation elements of this correlation matrix.

5.2 Results

The following sections present simulation results for the toy data set. First an example for separations of the toy data set is shown to give an impression of the quality of separation achieved by different algorithms. Furthermore, the Reconstruction Error for all of the sources, mixtures, and estimated sources is provided, which helps to interpret the plots in later sections. A discussion of the shifts used for the simulations is provided in the following section.

The plots in the next sections are grouped to illuminate four issues: The first is a comparison of the single shift heuristic with the average and optimal single shift, giving an impression of its usefulness (section 5.2.2). Then, second, a comparison of the heuristical single shift algorithm with the multiple shift algorithms, using standard sphering, is done (section 5.2.3). Following, as the third set of plots, is a comparison of standard sphering and noise-robust sphering techniques for the multiple shift algorithms (section 5.2.3). The fourth and final set visualizes the performance of multiple shift algorithms on data with spatially correlated sensor noise (section 5.2.4).

For the evaluation of noise-robustness simulations were performed with varying levels of sensor noise. Most graphs show plots of the Reconstruction Error against the noise level, measured in dB. Higher noise levels correspond to a lower decibel value and are thus visible in the left part of the plots, while lower noise levels appear in the right part. As the signal to noise ratio for zero noise is infinity, the corresponding value is not show in the plots. It is generally not very different from the first noise level shown in the right part of the plots, anyway.

In the end of this section a plot is shown which relates the noise level of correlated noise with the percentage of successful runs. For white noise nearly all separations were successful, so their plots are not shown.

It is obvious from the RE versus noise level plots that different mixing matrices yield different quantitative results. Generally, it appears that for higher condition numbers of the matrices (greater than about 10 to 15) the resulting Reconstruction Error is less stable; the error bars in the plots of the RE are significantly higher. A separation is often still possible, but it depends more on the actual noise (not only its variance) and for the gradient descent algorithm also on the initial parameters. A general law, that mixing matrices with high condition numbers are harder to invert

[2] See explanation in section 4.2.3.

than for low condition numbers, was not observable; even for condition numbers of above 20 it was sometimes still possible for algorithms to find good source estimates.

Results were obtained for two different mixing matrices, which can demonstrate the typical behavior of the algorithms. The matrices are given in table 5.1.

Matrix 1 condition number 8.57	Matrix 2 condition number 3.73
$\begin{pmatrix} -0.9497 & -1.6834 & -1.4192 \\ 1.0313 & -1.6144 & -1.6555 \\ 1.5354 & 0.5658 & 1.1511 \end{pmatrix}$	$\begin{pmatrix} -0.4326 & 0.2877 & 1.1892 \\ -1.6656 & -1.1465 & -0.0376 \\ 0.1253 & 1.1909 & 0.3273 \end{pmatrix}$

Table 5.1: Two mixing matrices used for noise analysis experiments.

For each mixing matrix and noise level several (usually 10) source separations are performed, where each time different noise (of the same variance) is added to the mixtures. The dpa algorithms are run 3 times on each mixture set, with different initialization; thereafter, the best run is counted. This is intended to compensate for the dependence on initial values of the parameters of the gradient descent optimization. All other algorithms are deterministic for a given mixture and do not need multiple runs.

The variance of the normally 10 runs for each algorithm and noise level is shown in the plots as an error bar at 2 $\times$ SEM (Standard Error of the Mean) above and below the mean. The SEM is the standard deviation σ of the Reconstruction Error for the runs normalized by the square root of the number of runs n:

$$\text{SEM} = \frac{\sigma}{\sqrt{n}} \tag{5.2}$$

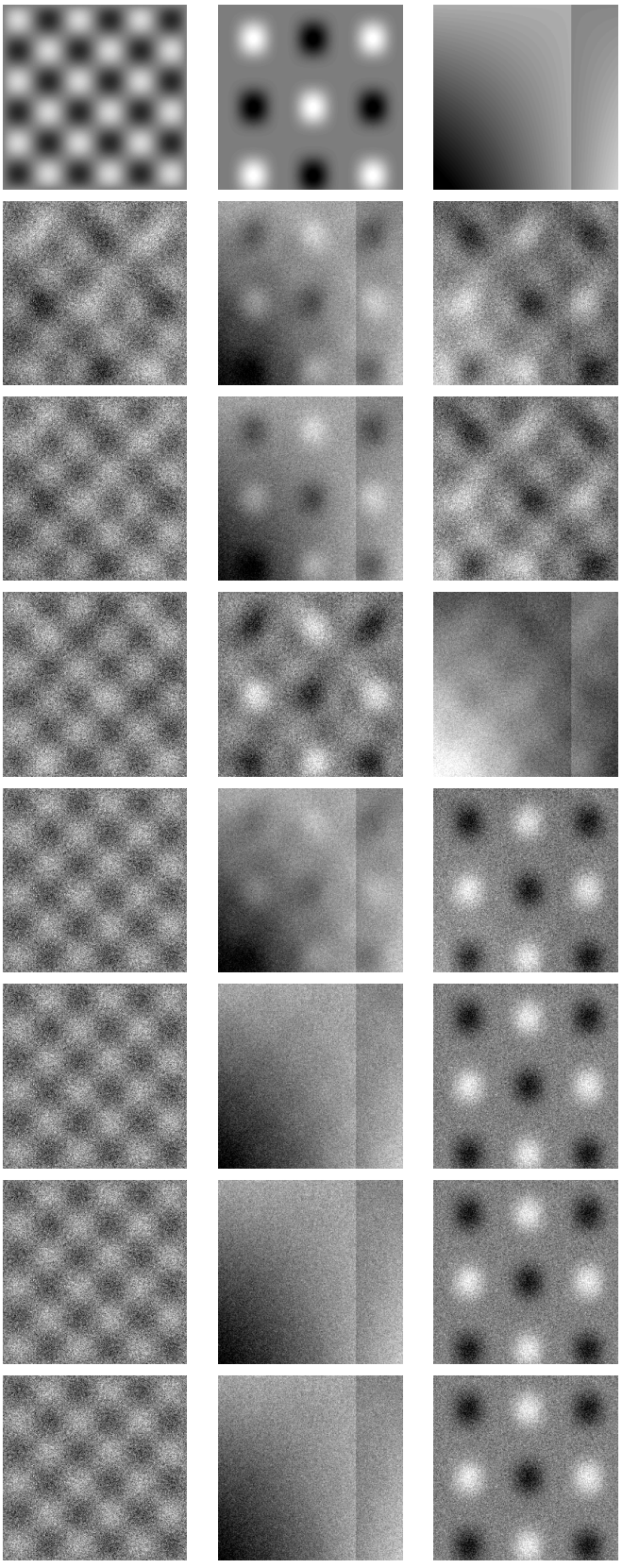
5.2.1 Separation Example

To give an impression of the separation capabilities of different algorithms some separation results are presented in figure 5.3. It is obvious that the sensor noise cannot be filtered out by these BSS algorithms (all estimated sources are still grainy), but the sources are clearly recognizable. The cor and to a lesser degree the jac0 algorithm let traces of one source be visible in another one. The other algorithms do a very good separation. To give an impression of the meaning of the Reconstruction Error, it is given next to each separation result.

The mixing matrix (condition number 5.1) used for this example is:

$$\begin{pmatrix} -0.3497 & 0.4216 & 0.1838 \\ 0.1915 & -0.9357 & 1.9059 \\ -0.2875 & -0.6827 & -0.6122 \end{pmatrix}$$

Another important thing to note is that the RE for the jac1 and dpa1 algorithms is less than for the original sources, which is larger than zero. This means that the original sources are not completely



Sources: Original sources used for the simulations
RE = 0.0236

Mixtures: Mixtures with added sensor noise (std. dev. 0.8) condition number of mixing matrix: 5.1
RE = Inf

Sphered Data:
The data after sphering using a shift by one pixel
RE = 0.4496

cor: Separation result of the cor algorithm
RE = 0.2163

jac0: Separation result of the jac0 algorithm
RE = 0.1359

dpa0: Separation result of the dpa0 algorithm
RE = 0.0301

jac1: Separation result of the jac1 algorithm
RE = 0.0176

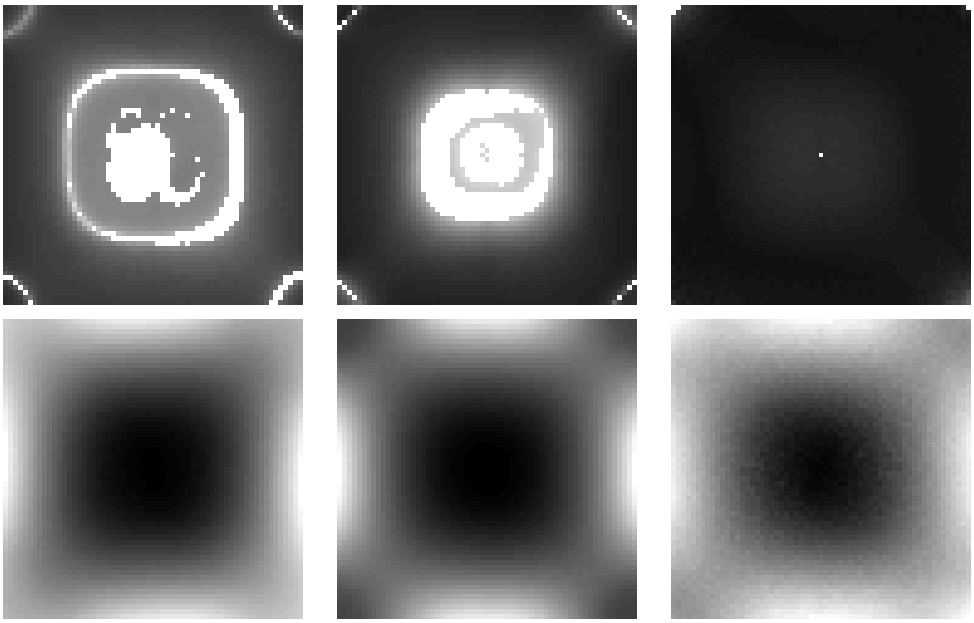
dpa1: Separation result of the dpa1 algorithm
RE = 0.0191

Figure 5.3: Example of separations and the Reconstruction Error achieved by different algorithms.

uncorrelated, and the two mentioned algorithms succeed in finding estimated sources which have less correlation than the original ones.

5.2.2 Choice of Shifts

Experience shows that for the ESD algorithm the choice of the *single shift* is very critical for the quality of separation. To automate the selection of the shift an algorithm (cor) was implemented, which uses the sizes of the cross-correlation matrices of the respective shifts for its choice; figure 5.4 gives in the second row three examples of how this heuristic looks like for different shifts. Compared with the images showing the RE for the corresponding shifts (these are shown in the top row), it is perceptible that generally the regions with maximal values (light) for the heuristic are in places where the separation quality is good (dark value). In the rightmost images a pathological example is shown, where the maximum of the heuristic is just at one of the few shifts for which the separation fails.



maximum 1.4134
condition number 8.57

maximum 0.6028
condition number 3.73

maximum 0.1735
condition number 6.18

Figure 5.4: This figure shows the quality of separation for different shifts (using the single-shift ESD algorithm) and the value of the correlation heuristic used for the cor algorithm for mixtures of three different mixing matrices. In the top row the RE values are shown as gray levels normalized between 0.0 and 1.0. I.e. a white pixel denotes a failure, while dark pixels indicate a good separation for the corresponding shift. The correlation heuristic images in the bottom row are normalized between 0.0 and the maximum of the values for each matrix, which is printed below each image. The zero shift is in the middle of the images and the borders correspond to shifts of 30 pixels up, down, left and right.

Figure 5.5 shows, that the heuristical choice normally yields a quality much better than for the average shift (at least for this data set) and relatively close to the optimal single shift. For all evaluated noise levels the RE value is very close to or at least closer to the optimal shift value than to the average one. For some mixing matrices, on the other hand, its choice of shift can

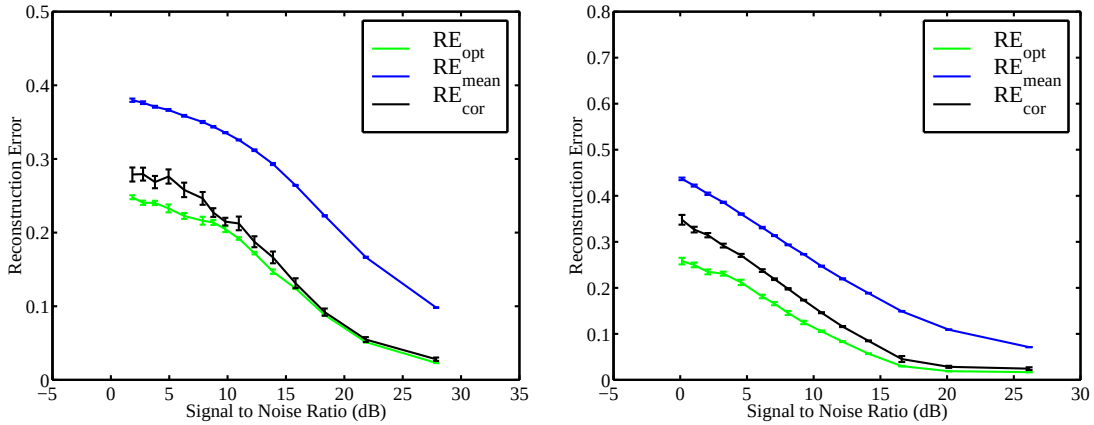


Figure 5.5: Shift selection strategies are compared for single-shift ESD algorithms. The cor algorithm often selects shifts which yield estimated sources whose RE is much better than for the average shift, for low noise levels very close to that of the optimal single shift.

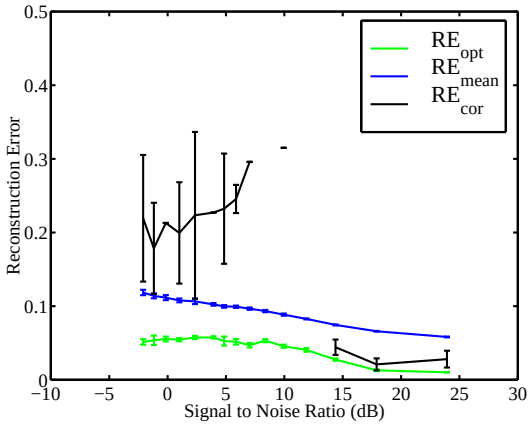


Figure 5.6: An example of results for a mixture matrix (condition number 6.18), which makes the cor algorithm to select bad shifts.

be bad, as the example in figure 5.6 shows. In this case the separation for medium noise levels fail completely, and for higher noise levels it fails often, while the rest show a high SEM with a mean above the average shift. Only for low noise levels till about 15 dB the heuristic works well. Unfortunately the correlation heuristic has its maximum (in the evaluated region of 30 shifts in each direction) often at one of the few points where separation fails (see figure 5.4). For the other mixing matrices it has its maximum in the regions with very good separation qualities.

Personal experience with the *multi-shift* algorithms indicates that the exact choice for the set of shifts is not critical, although it is useful to take the range of auto- and cross-correlations of the sources into account. Experiments have shown that a map of the separation quality (RE) for different shifts is relatively smooth, i.e. similar shifts mostly give comparable separation results (see top row of figure 5.4).

For the simulations with the artificial data in this chapter a set of shifts in form of a star was used

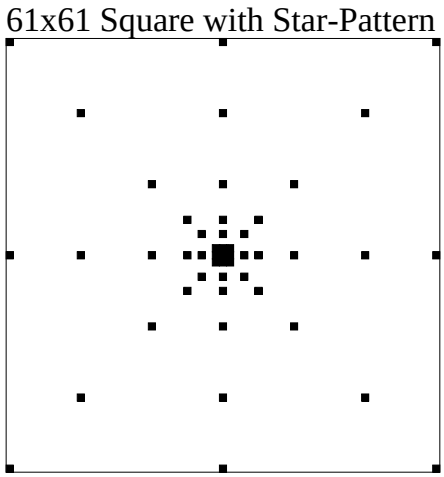


Figure 5.7: The set of shifts used for multi-shift simulations with toy data.

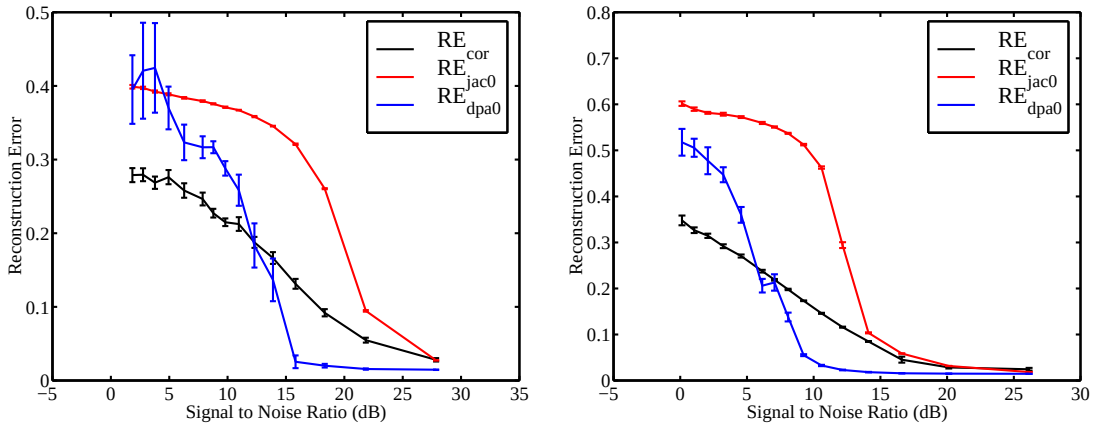


Figure 5.8: Single- and multi-shift algorithms are compared for two different mixing matrices. Only standard sphering is used.

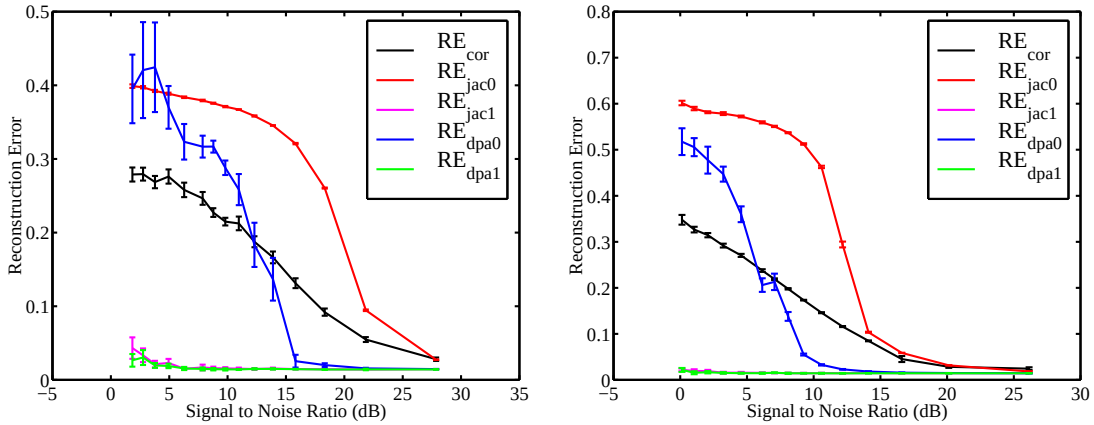
for all algorithm. it is shown in figure 5.7. The star includes all 8 shifts with a distance of 1 pixel, as well as shifts of 3, 5, 10, 20, or 30 pixels up, down, left, right, and in direction of the 4 diagonals.

5.2.3 Sensitivity to White Sensor Noise

Nearly all simulation runs in this subsection were successful, i.e. each algorithm returned sources with a finite Reconstruction Error for almost every mixture. Only two runs of the dpa0 algorithm on the mixing matrix with condition number 8.57 were unsuccessful.

Comparison Single/Multiple Shifts

In figure 5.8 the Reconstruction Error is plotted for a single- and two multiple-shift algorithms, for two different mixing matrices. Only the standard sphering technique is used. The results



condition number 8.57 condition number 3.73

Figure 5.9: The separation performance of multi-shift algorithms; standard sphering and noise-robust sphering are compared.

are typical: The Jacobi algorithm has problems already for very low noise levels and usually returns the worst source estimations. The gradient descent algorithm is performs very well for low to medium noise levels. It is often worse than the single-shift algorithm (cor) for high noise levels, but that depends on the mixing matrix; particularly for mixing matrices with high condition numbers, above 10, no prediction about the order of the gradient descent and the single shift algorithms is possible. Furthermore, the error bars often become very large for the gradient descent algorithm for high noise levels.

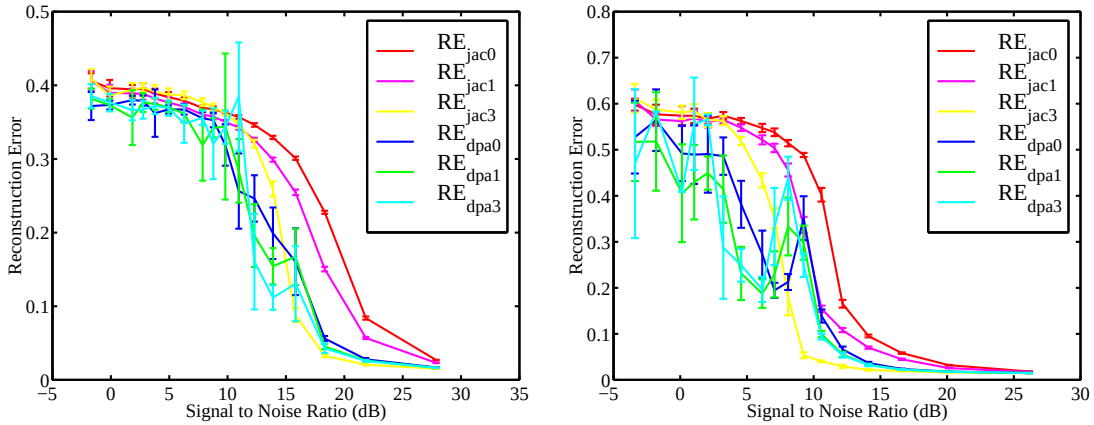
An important observation, when figure 5.8 is compared with figure 5.5), is that the gradient descent method can return source estimates, which are better than those of the optimal single shift. Although this is not true for all mixing matrices it shows that it can be advantageous to use information about correlations for several shifts.

It is obvious from the bad performance of the Jacobi algorithm with standard sphering (jac0) that it is unable to cope effectively with sensor noise. For almost all mixing matrices its performance is much worse than that of all other algorithms.

Noise-Robust Sphering

The plots in figure 5.9 show the performance of both multi-shift algorithms (dpa and jac) using standard sphering, as well as using noise robust sphering. As a reference the Reconstruction Error for the single shift algorithm is also given. While the standard sphering variants have the same curves as in figure 5.8, the variants using noise robust sphering provide very good source estimates for all noise levels. For high noise levels these are as good as for low noise levels, and very much better than those given by the single-shift algorithm or the standard sphering variants. The separation quality of jac1 and dpa1 is not significantly different.

It is obvious that the separation performance greatly improves by using the noise robust sphering technique. This observation holds for all mixing matrices analyzed by me, and although the Reconstruction Error for both algorithms often increases for higher noise levels, if other mixing matrices are used, it normally stays far below that of the cor, dpa0 and jac0 algorithms.



condition number 8.57 condition number 3.73

Figure 5.10: The separation performance of multi-shift algorithms evaluated for spatially correlated noise; standard sphering and noise-robust sphering for different shifts are compared.

Another point is noteworthy: The jac algorithm is more sensitive to sphering errors than the dpa algorithm. The gradient descent algorithm using the standard sphering technique is able to compensate for sphering errors, and thus shows reasonable performance at least in the medium noise level zone. The Jacobi algorithm gains much more by using the noise robust sphering, because it can not compensate for incorrect sphering.

5.2.4 Sensitivity to Noise with Non-Zero Spatial Auto-Correlation

During analysis of the data set presented in section 6 the question arose how BSS algorithms would perform if the sensor noise was spatially correlated. The possibility of the noise in the optical imaging data being correlated could not be excluded by theoretical considerations, and so the effect of correlated noise was evaluated for the toy data set.

Spatially correlated noise can influence the separation performance of the algorithms presented in this thesis, because they use information in spatial correlations of the mixtures, which is modified by correlated noise.

$$\begin{aligned} \mathbf{C}(\Delta\mathbf{r}) &= \left\langle \mathbf{y}'(\mathbf{r})\mathbf{y}'^{T}(\mathbf{r}+\Delta\mathbf{r}) \right\rangle_{\mathbf{r}} \\ &= \left\langle \mathbf{DAs}(\mathbf{r})\mathbf{s}^{T}(\mathbf{r}+\Delta\mathbf{r})\mathbf{A}^{T}\mathbf{D}^{T} \right\rangle_{\mathbf{r}} + \left\langle \mathbf{Dn}(\mathbf{r})\mathbf{n}^{T}(\mathbf{r}+\Delta\mathbf{r})\mathbf{D}^{T} \right\rangle_{\mathbf{r}} + \\ &\quad \left\langle \mathbf{DAs}(\mathbf{r})\mathbf{n}^{T}(\mathbf{r}+\Delta\mathbf{r})\mathbf{D}^{T} \right\rangle_{\mathbf{r}} + \left\langle \mathbf{Dn}(\mathbf{r})\mathbf{s}^{T}(\mathbf{r}+\Delta\mathbf{r})\mathbf{A}^{T}\mathbf{D}^{T} \right\rangle_{\mathbf{r}} \end{aligned} \tag{5.3}$$

If the added noise is spatially white then the expectation value $\left\langle \mathbf{Dn}(\mathbf{r})\mathbf{n}^{T}(\mathbf{r}+\Delta\mathbf{r})\mathbf{D}^{T} \right\rangle_{\mathbf{r}}$ is zero except for the zero shift $\Delta\mathbf{r} = [0,0]^{T}$. For noise which is independent of the sources, the third and fourth term (correlation between sources and shifted noise, and between noise and shifted sources) are also zero. Then the noise has no influence on correlations for others than the zero shift (for non-vanishing noise its variance always influences correlations for the zero shift). If, on the other

hand, the noise is correlated, then it influences all correlation matrices whose shifts are in the range where the auto-correlation of the noise is unequal to zero.

The noise used for the simulations described in this subsection is produced by filtering (blurring) Gaussian white noise with a Gaussian of variance 1.0 pixel and normalizing the result to the given noise level (standard deviation) by multiplying the noise vectors by the ratio of wanted standard deviation and actual standard deviation.

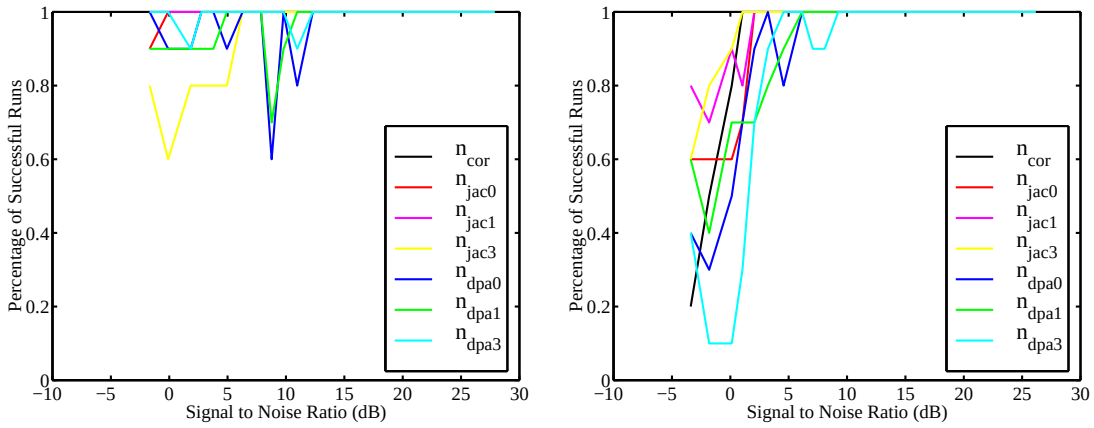
The simulations were performed using sphering matrices based on shifts between 0 and 6 pixels. 1 pixel, as used in jac1 and dpa1 are very few, as the correlation of the noise is still large for a shift of one pixel. The results are not shown for all shifts, as the separation quality does not improve significantly above shifts of 3 pixels, while the results become less stable for larger shifts (more unsuccessful separations). The number of approximately 3 pixels is plausible, because for values of three times the standard deviation from the mean a Gaussian is almost zero, i.e. the auto-correlation of the noise should be close to zero for a shift of 3 pixels.

Figure 5.10 shows the Reconstruction Errors of the multi-shift algorithms for different sphering shifts. The effect for the Jacobi algorithm is very clear, due to the small error bars: For a sphering shift of 0 pixels it performs worst, a shift of 1 pixel improves its performance, while the optimal performance is reached for a sphering shift of 3 pixels. Larger shifts do not show a significant improvement (not shown in the plots). For the gradient descent algorithm the effect is similar, although the differences for the various sphering shifts are not as significant, because of the larger error bars. Often the gradient descent method performs better than the Jacobi method, because its mean performance is often below that of the other; on the other hand, the error bars for the gradient descent algorithm are much larger, indicating less reliability of the separation quality. In the low noise region the jac3 algorithm often has light advantages over all dpa variants; in the regions with higher noise levels, the gradient descent algorithms give slightly better performance.

The results in these plots again show that the Jacobi method depends very much on the sphering technique used. The differences for the gradient descent variants are less significant and indicate that these are much better in adapting to incorrect sphering.

While the simulation runs for white noise were successful for nearly all algorithms and noise levels, more unsuccessful separation occurred for the simulations using spatially correlated noise, even for the same mixing matrices. Figure 5.11 shows the percentage of successful runs of the algorithms for different noise levels. For low to medium noise levels most simulation runs return successfully separated source estimates. For high noise levels (left part of the plots) the Jacobi algorithm variants have usually a higher number of successful runs than the gradient descent algorithm. The single shift algorithm usually returns less successfully separated source estimates than the Jacobi algorithm, but more than the gradient descent.

Up to a shift of 3 pixels in the sphering procedure the separation performance normally gets better with larger shifts (both for jac* and dpa* algorithms). A sphering with a shift of 3 pixels, on the other hand, already seems to be less reliable, despite its otherwise good separation quality, and often results in an increased number of unsuccessful runs. This is understandable, as the the auto-correlation of the noise becomes nearly zero for shifts of three, eliminating contamination of noise and improving the separation quality. On the other, the auto-correlations of the sources normally become more and more different from the variance for larger shifts, which making the sphering and consequently the Error minimization more erratic.



condition number 8.57 condition number 3.73

Figure 5.11: The success rates (percentage of runs where $\mathsf{RE} \neq \infty$) of the simulations in figure 5.10.

Chapter 6

Results on Optical Imaging Data

This chapter covers the simulations performed on optical imaging data sets. The data set used here is the same as described in [SSM+99], although a slightly different preprocessing is applied before the actual BSS is performed. In the following the format of the recorded frames is described, followed by an explanation of its preprocessing; results of source separation simulations are presented thereafter.

6.1 Optical Imaging Data Set

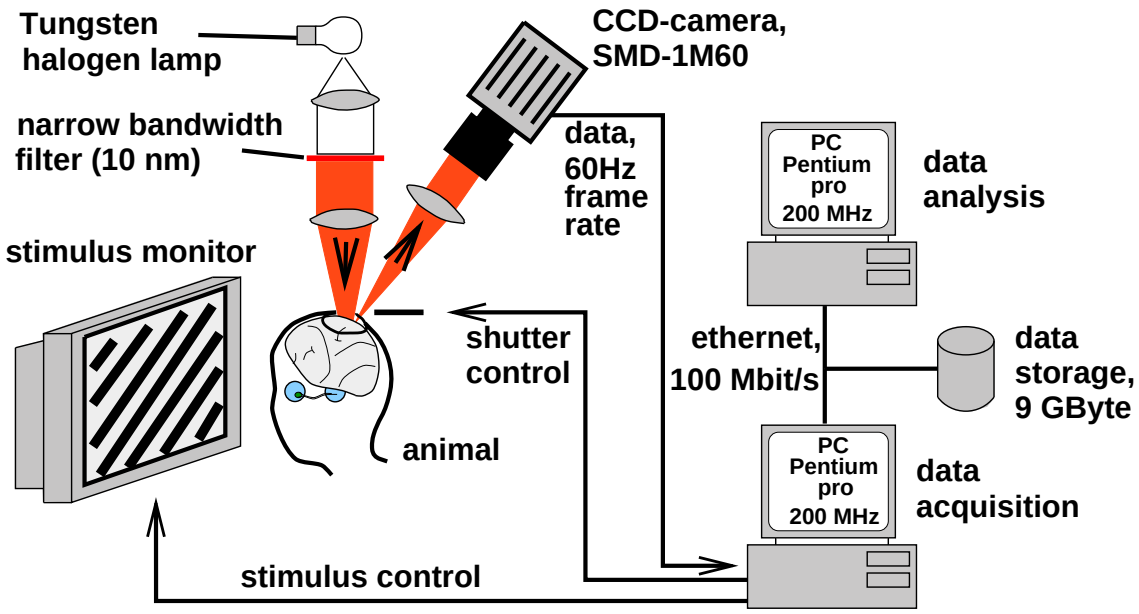


Figure 6.1: The experimental setup used for acquisition of the optical imaging data. The reflections of orange light (wavelength 633 nm) from a part of the visual cortex of a monkey are recorded by a CCD camera, possibly during presentation of visual stimuli to the monkey. The frame stream is preprocessed and stored for later analysis by a PC.

The experimental setup for the recording of the optical imaging data set is shown in figure 6.1.

The data is provided by the team of J. Lund, Department of Opthalmology, University College London. A CCD camera collects orange light (wavelength 633 nm) reflections from the primary visual cortex of a macaque monkey during presentation of stimuli. The structure of the recorded frames is visualized in figure 6.2. During each trial frames are recorded for 8 seconds, with a frame rate of 16 Hertz. During the first two seconds no stimulus is present; during this period blank image frames are obtained, which can be used for preprocessing later. Starting with time $t_1 = 2$ sec. a stimulus is shown for a duration of four seconds (ending at $t_2 = 6$ sec.). After the stimulus ends the recording is continued for another 2 seconds. Between two successive trials a recovery period of 15 seconds passes. Consecutive trials are usually performed for different stimulus conditions, to exclude long term reactions in the visual cortex due to repeated stimulation with the same stimulus.

The frames recorded by the camera consist of 1024×1024 pixels each. From these 256×256 pixel frames are extracted for further analysis. The width of one pixel in these frames corresponds to 14.8 ± 0.5 μm of cortical tissue. The 256 pixel frame width accordingly correspond to 3.7 ± 0.1 mm of the cortex.

All simulations in this chapter used image stacks, which were recorded with stimuli applied to either only the right eye or only the left eye. These stimuli were intended to produce frame stacks which could be used for extraction of ocular dominance maps. Such maps indicate, which regions react to stimulation of the left eye and which to that of the right eye.

6.2 Preprocessing of Data

The recorded frames described in the previous section contain a large amount of noise and artifacts which are unrelated to neural activity in the cortex. To make the extraction of the mapping signal from these frames easier for the source separation algorithms some preprocessing steps are applied, which generally try to improve the signal to noise ratio in the frame stack.

One method used is averaging over trials. Several (here 16) trials for each stimulus condition are recorded, which are later summed up to average over noise. Frames corresponding to the same point of time (with respect to the stimulus onset) in different trials are added to give a frame of an averaged frame stack. The mapping signal, whose extraction is the goal of these experiments, and the global signal should have approximately the same time series in all trials and are thus amplified, while noise which is unrelated to the stimulus will show different time series and is thus partially canceled out.

Another approach for improving the signal to noise ratio is temporal averaging. Each of the images in the data set processed by the BSS algorithms is the sum of 15 frames recorded by the camera; thus each image represents the frames recorded during one second. The SNR of the mapping signal (and again the global signal) versus fast changing noise, which is different between frames recorded closely after another will be improved using this procedure. Fast changing noise is e.g. photon shot noise.

The first image of the eight resulting ones, which is recorded while no stimulus is present, is subtracted from all other images. This process is called first frame analysis. It is intended to remove artifacts with very slow temporal changes, which are approximately constant during the 8 seconds of recording time for each trial. These could be e.g. blood vessels. The problem with

blood vessels is that they change their size due to changes in blood flow (stimulus related changes as well as not related ones) during the experiment; thus this approach is not always successful in removing such artifacts. Furthermore, movements of the recorded cortex with respect to the camera can make first frame analysis problematic. Such movements could for example occur if the optical chamber is not sufficiently tightened on the skull, if the head of the animal is not fixed, or because of heartbeat or respiration. After first frame analysis the first image is completely zero and left out in the presentations of results; these show 7 images, corresponding to seconds 1 to 7, instead of 0 to 8.

In the experiment presented in the following sections a large vessel occupies part of the image (see arrow in figure 6.3). It is masked out in the images by setting all pixels in its range to zero; furthermore this region is ignored for calculations of means and variances in the algorithms. The masked part of the images is visible in the images presented later as an area with a uniform gray value. Additionally to the large vessel artifact a peak in reflectance approximately in the middle of the images is masked out. This peak could be caused by direct reflection of the light source on the surface of the cortex.

To further improve the signal to noise ratio in the images I investigated lowpass filtering with different frequency limits. This procedure eliminates components with a spatial frequency above a given number of cycles per 256 pixels (the image width). The frequency limit which visually gave the best results on the optical imaging data was 50 cycles, and lowpass filtering with this limit was used in most of the presented simulations. Alternatively binning of neighboring pixels could be used to spatially average over noise. This would make the images smaller and thus the algorithms faster. On the other hand, time is generally not very critical in these experiments, and the use of filtering allows a finer control of the averaging by changing the number of cycles.

Finally the Blind Source Separation algorithms are applied on two types of image stacks: Single condition stacks and difference stacks. In the difference stack the quality of the mapping signal is further improved. Subtraction of maps with orthogonal stimuli amplifies the part of the mapping signal, which changes between the presented stimuli. Biological noise, especially the global signal and vessel artifacts which are still remaining after the preceding steps, is reduced greatly. On the other hand this step also amplifies sensor noise like photon shot noise, and the interpretation of the resulting maps is different for difference maps than for single condition maps, because of the underlying assumptions about orthogonal stimuli. For some types of stimuli it may also be hard to come up with orthogonal ones.

The following is a summary of all steps used to improve the signal to noise ratio for the mapping signal compared to the raw frames recorded by the camera:

- Summation of several trials for the same stimulus
- Temporal summation (each image contains frames of one second)
- Binning of pixels
- First frame analysis
- Masking of areas with too much contamination by noise or biological artifacts
- Lowpass filtering
- Difference stack

- The BSS algorithm was optimized for its noise robustness (multiple shifts, non-rotational demixing matrix)

6.3 ESD for Optical Imaging

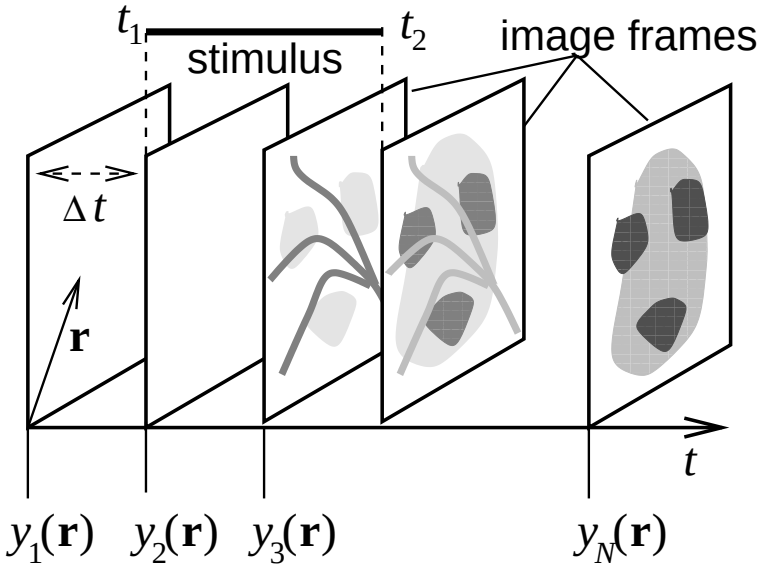


Figure 6.2: Format of recorded image stacks. The first frame is blank (no stimulus is present); from $t_1 = 2$ sec. to $t_2 = 6$ sec. the stimulus is presented. The total number of frames recorded is 120, these are later reduced to 7 by temporal averaging and first frame analysis. Each of the final images is the average over all frames the camera recorded during one second (Δt is 1 second). Features gradually pop up and vanish at characteristic times. $y_i(\mathbf{r})$ is the value of the pixel at location $\mathbf{r}$ in image number i. Taken from [SSM$^+$99].

This section gives on overview about how the BSS framework is applied to the optical imaging data set, and how its structure can be interpreted. The organization of the image stack is sketched in figure 6.2. Certain prototype images can be distinguished, in the sketch these are vessels, background, and the mapping signal. Prototypes are modeled to be linearly mixed with different coefficients in each image. The spatial prototype patterns are what is called sources in the BSS framework (see section 4.1); they could represent vessel patterns, the global signals, biological noise, and the mapping signal (local activity of neurons related to stimulus). Thus the mixing coefficients tell how strongly each prototype pattern is present in different frames and can be interpreted as the time series of the respective patterns. The assumption of linear mixtures is necessary for the application of the BSS algorithms presented in section 4. The recorded images also contain much sensor noise, which cannot be modeled as a separate source.

The image stacks used in the optical imaging experiments are shown in figure 6.3. The top row shows the seven images of the single condition stack, the bottom row those of the difference stack. Both are shown without lowpass filtering and without masking, but after first frame analysis. In the top row, the mapping signal is invisible; only the changes of reflectance from the vessels and a general change in background intensity are visible. The large vessel (marked by the arrow in the second frame of the top row), which is masked out for the simulations, is very prominent in the

single condition stack. The intensity of reflection by this vessel, and also by the background, is highest during the period the stimulus was presented (for the images 2 to 5). In the difference stack the vessel artifacts are almost removed. In the first image, which was recorded while no stimulus was present, one of its branches still pops up, but otherwise the vessels are no longer visible. Instead, from the fourth to the seventh image the mapping signal (in form of ocular dominance stripes) is visible.

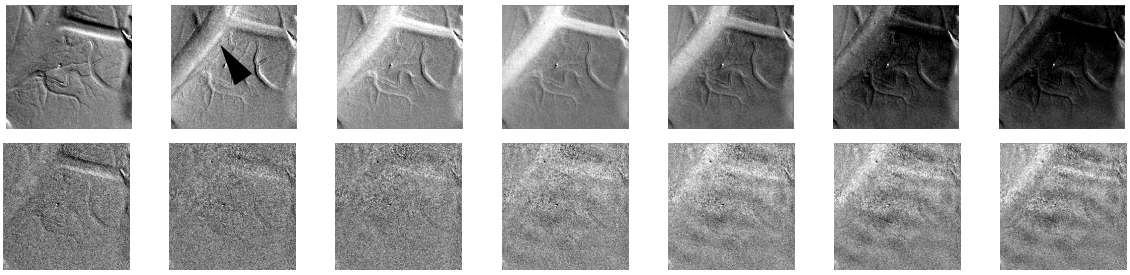


Figure 6.3: The optical imaging stacks used for Blind Source Separation. The stacks are shown without masking and without lowpass filtering, but after first frame analysis. The top row shows the time-series for the single condition stack (primary visual cortex, ocular dominance experiment, left eye), the bottom images the one of the difference stack. The stimulus was presented during recording of images 2, 3, 4, and 5 in each row. In the second image of the top row a large vessel is marked by an arrow; this vessel was masked out for most experiments.

The algorithms in this thesis try to recover the prototype images from the mixtures (recorded and preprocessed images) by using their different spatial structure, i.e. different spatial autocorrelation structures of the prototypes, and the assumption of zero cross-correlation function. The results in section 6.4 show that in most cases the spatial autocorrelation structure of the mapping signal is different enough from that of other prototypes for successful separation.

An alternative would have been to use different temporal structures (time series) of the prototypes. But the temporal structure of the signal specific to local activity of neurons is related to the global signal, and so it may be hard to use temporal autocorrelation structures for the separation task. It was tried by others, but the result turned out to be not very promising and was not further pursued. Further comments concerning this topic are given in section 4.1.2.

6.3.1 Statistical Characterization of Optical Imaging Data

Noise

The experiments on artificial data were mainly intended to gain information about the behavior of the presented algorithms for different levels and types of noise, and sphering techniques. The optical imaging data set was not suitable for this task, as neither sources nor noise (type and level) are known. Nevertheless some analyses on this set were done to get rough estimates of the noise present in the data, and to decide whether other analyses of noise had to be done on the artificial data set.

These analyses were done for the ocular dominance dominance experiment presented in the last section. The unfiltered frame stack with 120 frames, saved before the accumulation in one-second

intervals took place, was used. The first 16 frames of this stack (about the first second of the summed trials) were used for noise analysis, because they do not contain signal components related to the stimulus presentation (the stimulus presentation was started after the 32nd frame, after two seconds). All differences between successive images were taken as an estimate of the noise. It has to be noted that part of this noise is averaged out in the image stacks used for the BSS experiments, because of the accumulation of one second intervals.

One type of analysis concerns spatial correlations of the noise, i.e. spatial whiteness. First the mean of each difference frame was normalized to zero and its variance to one. Then the spatial auto-correlation was computed by shifting each frame by shift vectors in the square from $(-5,-5)$ to $(5,5)$ and multiplying it element-wise with its unshifted version (ignoring the pixels, which were beyond the border of the other frame). All multiplied pixels were summed and then the sum was normalized by the number of multiplied pixels. Thus an auto-correlation was obtained for every difference frame for shifts around the zero shift. These calculations showed no significant spatial auto-correlation for the noise. With the variance normalized to 1.0, all auto-correlations for non-zero shifts are below 0.058 for the single condition stack and below 0.02 for the difference stack. Thus both the single condition stack and the difference stack contain spatially white noise. The similar results for different sphering methods, for the gradient descent as well as for the Jacobi method (see section 6.4), also indicate that the spatial correlation of the noise is not critical in these data sets.

Analyses of the spatial mean of these difference series show an oscillation of about 2 Hertz (see figure 6.4). This is probably caused by pulses of the blood flow or respiration, and not intrinsic in the noise itself. Furthermore it should not influence the experiments on the 7-frame stacks, as those work with frames accumulated over 1 second, which approximately cancels out this oscillation.

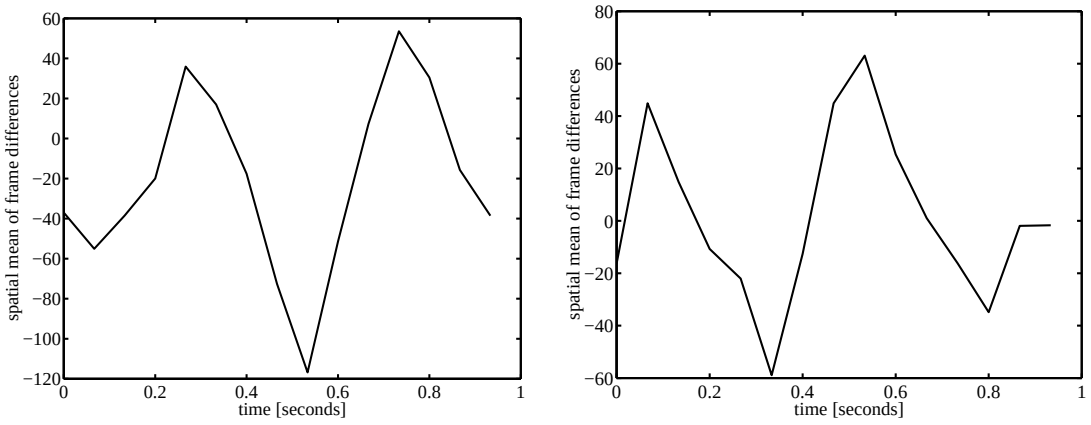


Figure 6.4: Plots of the spatial averages of the difference series of the first 16 frames (for the 120 frame stack). This corresponds to about the first second of recording, where no stimulus was applied. The left plot corresponds to the single condition stack, the right one to the difference stack. An oscillation with a frequency of about 2 Hertz is visible in both plots.

In figure 6.5 the temporal correlations of the frame difference series for different time lags are given. The procedure used for their computation is this formula, where $C_t(\Delta t)$ is the mean (over pixels) of the temporal correlation for time lag Δt, and $D_t(\mathbf{r})$ is pixel $\mathbf{r}$ of the difference between frames t and $t+1$ ($\mathbf{D}_t$ is image t of the difference series):

$$C_t(\Delta t) = \langle \langle D_t(\mathbf{r}) D_{t+\Delta t}(\mathbf{r}) \rangle_t \rangle_{\mathbf{r}} \tag{6.1}$$

In this equation the time index t runs from 1 to the number of difference images (15) minus Δt. Before application of this equation the mean of the time series for each pixel in the difference frame images are normalized to 0 and the variances to 1. This equation then computes the the temporal correlations of each pixel for a time lag of 0, $\frac{1}{15}$, $\frac{2}{15}$, ... seconds. Of the resulting images, showing the temporal correlations of all pixels for a given time lag, the spatial mean is taken. This results in a vector of 15 numbers, the spatial average of temporal correlations for 15 different time lags. These are shown in the plot.

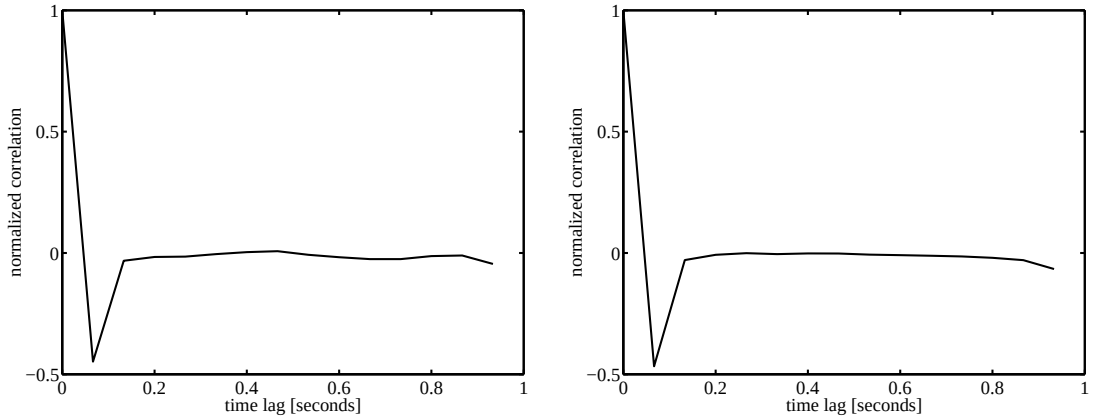


Figure 6.5: Two plots showing the temporal correlations of the difference series of the first 16 neighboring frames for the 120 frames stack for different time lags. The left plot shows the correlations for the single condition stack, the right for the difference stack.

For the single condition stack as well as for the difference stack the noise shows a strong negative correlation for a time lag of $\frac{1}{15}$ second, i.e. between neighboring frames in the 12-frame stack. Although I have no explanation for this phenomenon, its influence on BSS should not be too large, as the correlations are of very short duration ($\frac{1}{15}$ sec.), while the BSS algorithms work on images summed up over longer periods (1 sec.).

Concluding, it seems that, as far as these analyses went, the noise should not pose harder problems to the BSS algorithm than was tested on the artificial data set. The main point is that no spatial correlation seems to be present in the noise, which is beneficial especially for the Jacobi algorithm, but to a lesser degree also for the other algorithms.

Auto-correlation of Estimated Sources

The BSS experiments with the gradient descent algorithm seemed to provide very good source estimates for the optical imaging data set, both for difference stacks and for single condition stacks (although the maps obtained from latter can naturally not be as clear as the one obtained from the former). I used these results, which are given in section 6.4, to compute the auto-correlations of the estimated sources. The computed auto-correlations show some artifacts in form of circles, which are due to the lowpass filtering used. Nevertheless, they illustrate the differences of the autocorrelation structures for different (estimated) sources. The one of the mapping signal (image

4 for the single condition stack and image 6 for the difference stack) is much broader than all else, i.e. light color, indicating high auto-correlation, extends further.

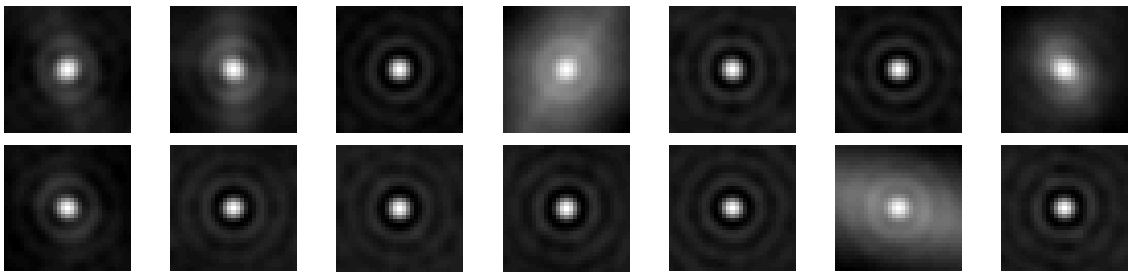


Figure 6.6: Auto-correlations of the sources estimated by the dpa1 algorithm, for the single condition stack (top) as well as for the difference stack (bottom). The zero-shift is in the middle, the borders correspond to shifts of 15 pixels up, down, left, and right. The images from left to right are in the same order as the corresponding sources given in figures 6.8 and 6.10.

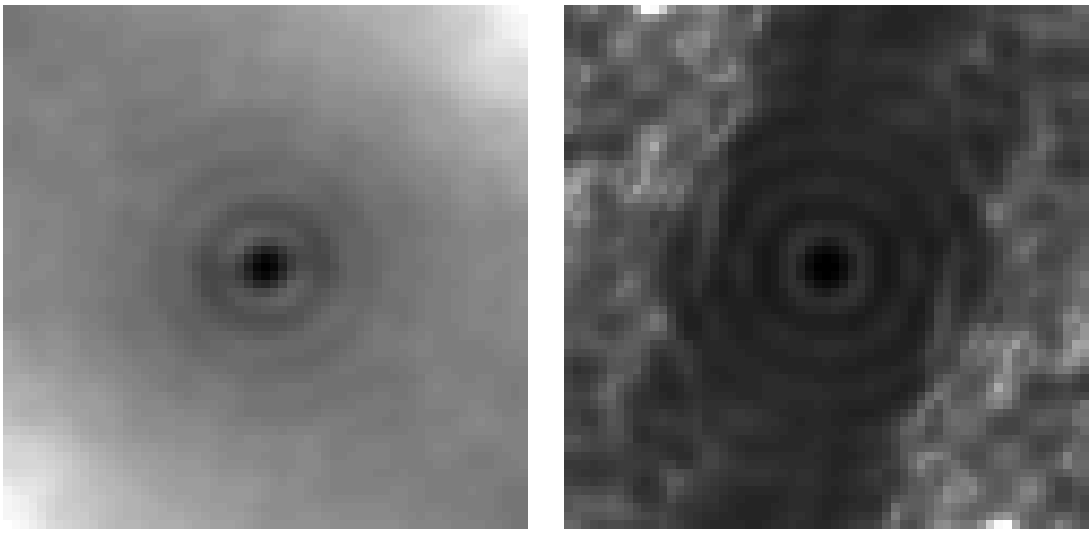
Correlation Heuristic

In the section about experiments on artificial data the cor algorithm seemed to perform reasonably well, compared to the optimal single shift. For the optical imaging data an objective separation quality measurement function is lacking (the RE cannot be used, because the real sources are unknown); to nonetheless be able to automate the choice of the shift used for separation this heuristic was devised. It would be unpracticable to explore, using visual inspection of the quality of the results, a wide range of possible shifts for the optical imaging data. Several shifts have to be analyzed to get a good chance that one is among them which gives a good separation.

Single shift experiments with arbitrarily chosen single shifts seemed to indicate that, for the optical imaging data set, the best shifts are in a relatively small region of about 5 pixels around the zero-shift. The cross-correlation heuristic, on the other hand, does not seem to choose shifts which give a good performance on the single condition stack. To explore this behavior further, I computed the value of the cross-correlation heuristic for all shifts in a square of 30 shifts into each direction (a square of 61×61 shifts).

In figure 6.7 the values of the cross-correlation heuristic for the sphered single condition and difference stacks are shown. The image of the heuristic values is relatively smooth for the single condition stack, while the difference stack yields one with more structure. Contrary to the observation, that good separation results are often achieved for small shifts, the values of the heuristic are very small for small shifts, and show several peaks (for the difference stack) or a general rise in their level (for the single condition stack) for larger shifts. The artifacts in form of circles around the zero shift are introduced by the lowpass filtering. I also computed the heuristic for unfiltered image stacks; they showed a less smooth structure, but with the same tendency to give high values for larger shifts. Especially the region, which shows several peaks in the given image for the difference stack (to the left and right of the “valley”), is very noisy (and gives very high single pixel peaks) when using unfiltered image stacks. The separation results corresponding to these high peaks are generally relatively poor.

It seems that the cross-correlation heuristic does not work as well for the optical imaging data as the experiments on artificial data suggested. This shows that this method cannot be an automated replacement for visual inspection for selection of single shifts.



Maximal value (white): 1.8318 Maximal value (white): 5.7583

Figure 6.7: Cross-correlation heuristic (see formula 4.15) for the sphered filtered single condition stack (left) and the sphered filtered difference stack (right). The middle pixel in each image corresponds to the zero shift, the borders to shifts of 30 pixels up, down, left, and right.

6.4 Results

The following two subsections show images which illustrate the separation ability of the dpa algorithm on optical imaging data. Different types of image stacks are used: Both single condition and difference stacks are analyzed. More examples are shown for single condition stacks, as it is more difficult to extract the ocular dominance maps for those. They are also important for interpretation of the maps, because they do not introduce the problem of selection of orthogonal stimuli, which is necessary to produce difference maps. Some assumptions have to be made about the organization of the cortex in order to select orthogonal stimuli: Disjunct neuron populations must be excited by those stimuli.

Besides the variation of the image stack types, the influence of preprocessing the stacks is evaluated and visualized by providing examples of separations using unfiltered and unmasked image stacks. To illustrate the benefits of the dpa algorithm, separation results of the other algorithms (cor and jac0) are also given.

In general experience shows that both the gradient descent and the Jacobi algorithm give good separation results on the optical imaging data. The Molgedey & Schuster algorithm is again very sensitive to the shift used for decorrelation. Even for good shifts (i.e. shifts giving a good ocular dominance map) it still fails most times in separating other prototype patterns like vessel artifacts as well as the multi-shift algorithms do. The Jacobi algorithm is in general very reliable. The

gradient descent algorithm, on the other hand, has to be run a few times, of which the best[1] result is chosen, to give reliable good results. But then it is often possible to obtain separations which are even slightly better (using visual inspection) than those given by the Jacobi algorithm.

6.4.1 Single Condition Maps

Figure 6.8 shows quite good separation results for a single condition stack, which was preprocessed using all of the techniques mentioned in section 6.2: Temporal and spatial averaging, averaging over trials, masking, and lowpass filtering. The seventh source estimate for the dpa0 algorithm and the fourth for the dpa1 algorithm are the ones representing the ocular dominance stripes; nearly nothing of this structure remains in the other source estimates, and almost no vessel artifacts are visible in the map. The projections of the maps onto the image stack also show a plausible time series, which begins around zero (no stimulus present in the beginning) and rises to maximum in the middle (when the stimulus ended) after which it slowly decays. None of the other sources shows a similar time series. The units of the Y-axis for the back-projections is arbitrary, as the sources can only be estimated up to an unknown scaling and permutation. The X-axis is the time in seconds from the start of recording of the trials. It ranges from 1 to 7 seconds because the first frame was used for first frame analysis. The stimulus was presented during seconds 2, 3, 4, and 5.

Most of the other estimated sources contain mainly blood vessel artifacts. Blood vessels probably have different spatial and temporal characteristics, depending on their size and distance to the main arteries, which supply larger areas with blood. So these artifacts are spread in a few sources. The rest of the estimated sources mainly show noise, which indicates that the number of mixtures, seven, is more than the number of sources present in the data. The fourth estimated source for the dpa0 algorithm could be interpreted as representing the global signal, caused by blood flow and volume changes, which spreads through the capillary bed from the large vessel. It is lighter close to the (masked) large vessel(s) on the top and left side and also along the smaller vessels. The farther away from the large and small vessels, the darker the gray value. Only one vessel artifact is not separated and still visible in this image.

Another observation which can be made in this figure is that both sphering variants for the gradient descent algorithm show similar performance. This also applies to the experiments I did with the two Jacobi algorithm variants. These observations indicate, that the noise level is not very critical in the image stacks. It seems, that the number of shifts used for decorrelation suffices for noise suppression. This conjecture is also supported by the sensitivity of the Molgedey & Schuster algorithm concerning the decorrelation shift.

Figure 6.9 shows the results of the cor and jac0 algorithms on the same single condition stack, to provide a possibility for comparison with the gradient descent algorithm in the previous figure. The jac1 algorithm is not shown because of its similar performance, with respect to the jac0 algorithms. In the separation result for the cor algorithm it is obvious that the oxygen level gradient is not separated from the ocular dominance map (image 5). Furthermore, vessel artifacts are visible in all of the estimated source and are not concentrated in a few images, as they were in the previous figure. The separation result of the Jacobi algorithm is nearly as good as that for the gradient descent method; only slight residuals of a vessel artifact are visible in the map (image 4). Otherwise the blood vessels are well concentrated in two sources, and the (supposed) oxygen gradient pattern is again visible as an own source (although this time inverted).

[1] As an objective measurement of the separation quality is not possible, because of the lack of a suitable cost function, visual inspection has to be used to determine the quality of separations.

6.4.2 Difference Maps

Figure 6.10 shows the separation results for the difference image stack (stack for stimulation of the right eye is subtracted from stack for the left eye), with otherwise same preprocessing as in the last subsection. The ocular dominance stripes are nicely separated in their own source, and also the back-projection on the image stack shows, as the only one, the expected time series. All other sources show almost only noise. This result is very similar for all analyzed algorithms, which is why only the result for the gradient descent algorithms is presented here.

6.4.3 Maps obtained for different preprocessing

In figure 6.11 the BSS results for image stacks using less preprocessing are given. The top is a single condition stack without the filtering used in previous experiments. Although the separation is quite well (vessel artifacts concentrated in few sources, the mapping signal concentrated in one source), the quality of the ocular dominance map is not as good as for the filtered stack. If the masking is left out for single condition stacks, the separation of the mapping signal fails completely (the mapping signal is not recognizable in any of the estimated sources). The correlations of the large vessel seem to be dominant in this case.

The bottom shows the data set and separation results for an unmasked and unfiltered difference stack. Due to the missing filtering the map quality is worse than visible in the previous figure; but still the separation works reasonably well (map concentrated in one source). Only the contamination by the large vessel not masked out this time mars the picture. For single condition images the contamination by the large vessel was too dominant for the acquisition of good ocular dominance maps from unmasked image stacks.

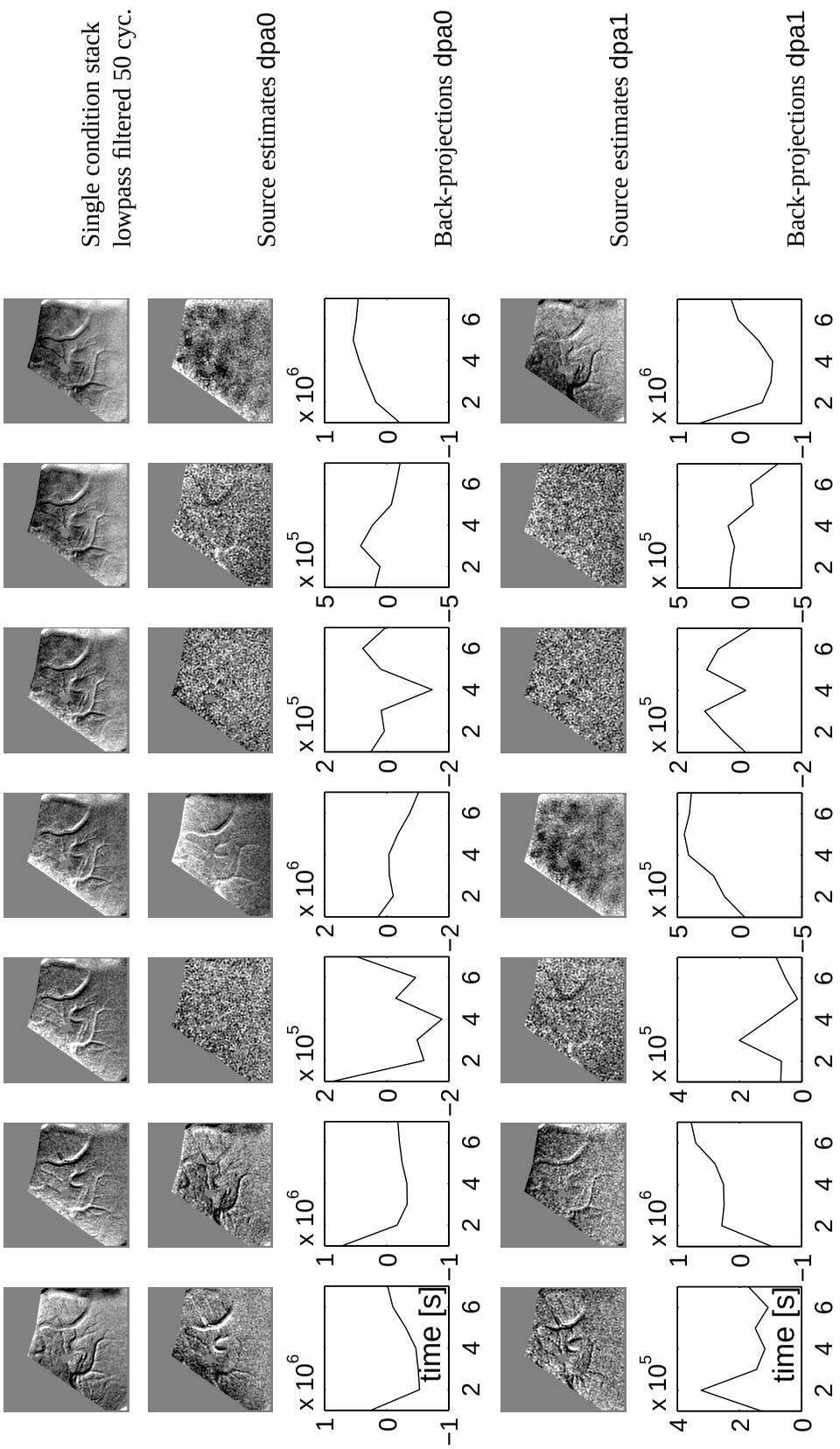


Figure 6.8: Blind Source Separation on single condition stack (ocular with lowpass filtering, 50 cycles. dpa0 and dpa1 algorithms.

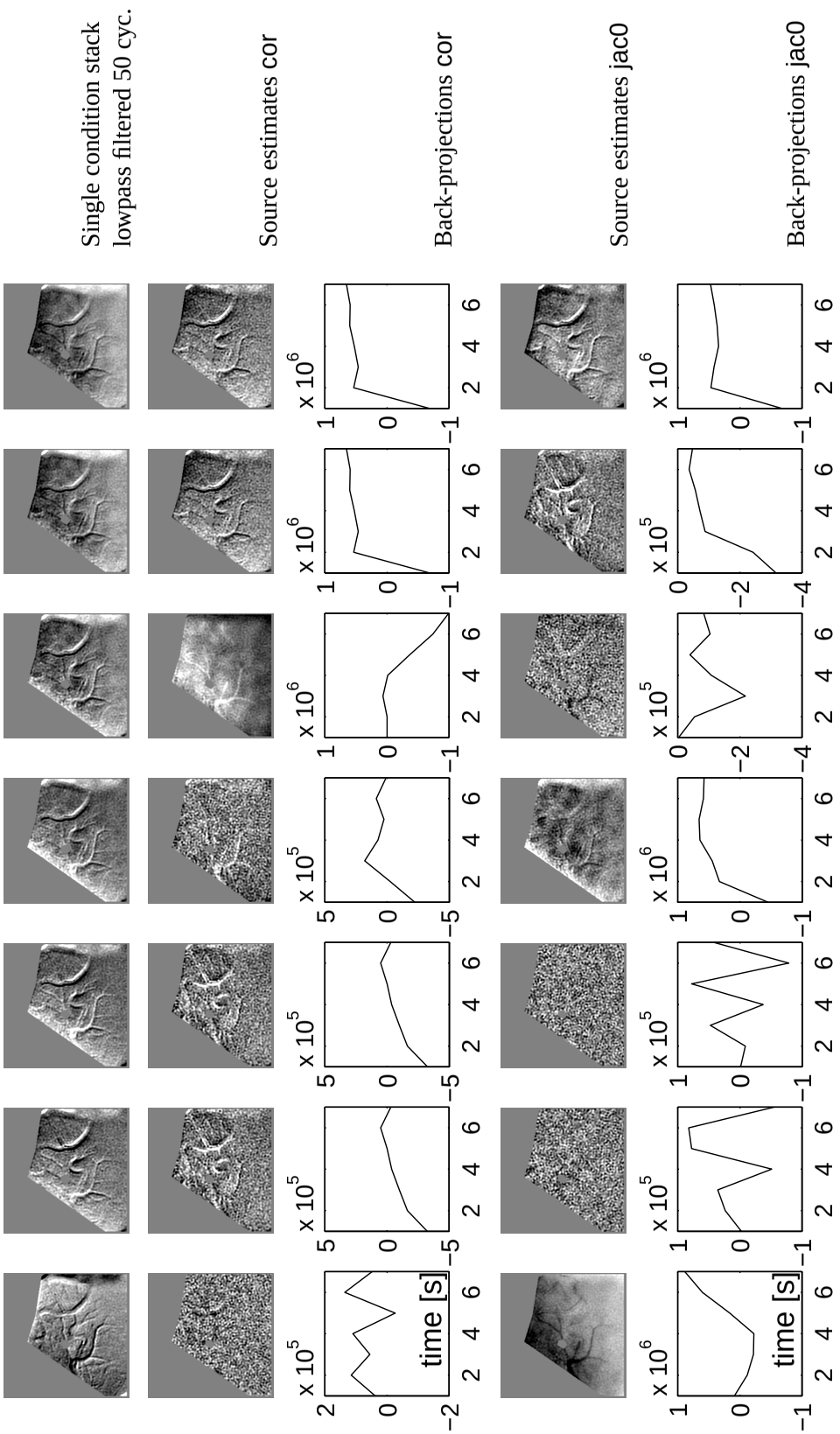


Figure 6.9: Blind Source Separation on single condition stack (ocular with lowpass filtering, 50 cycles. opt and jac0 algorithms.

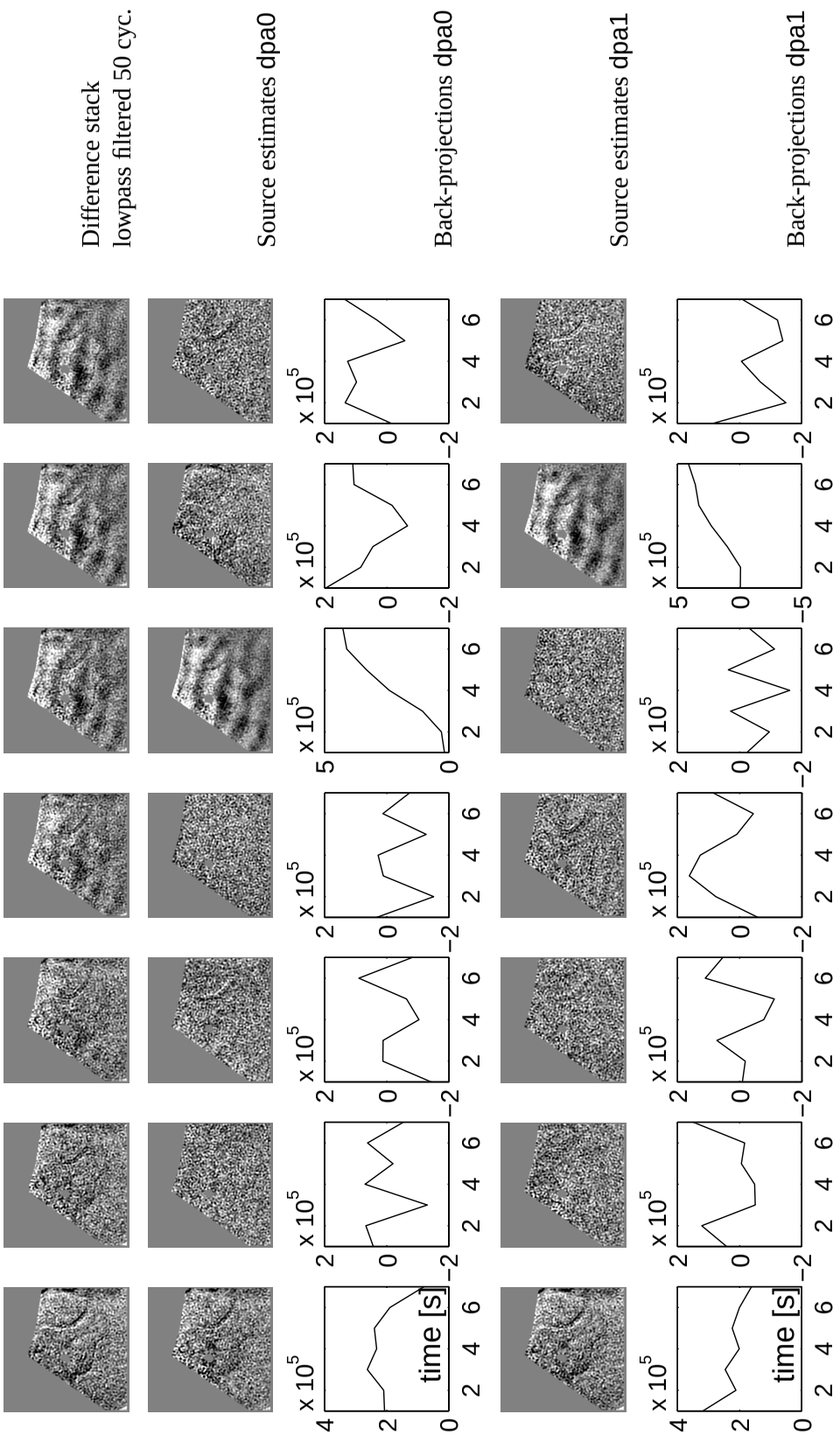


Figure 6.10: Blind Source Separation on difference stack with lowpass filtering, 50 cycles. `dpa0` and `dpa1` algorithms.

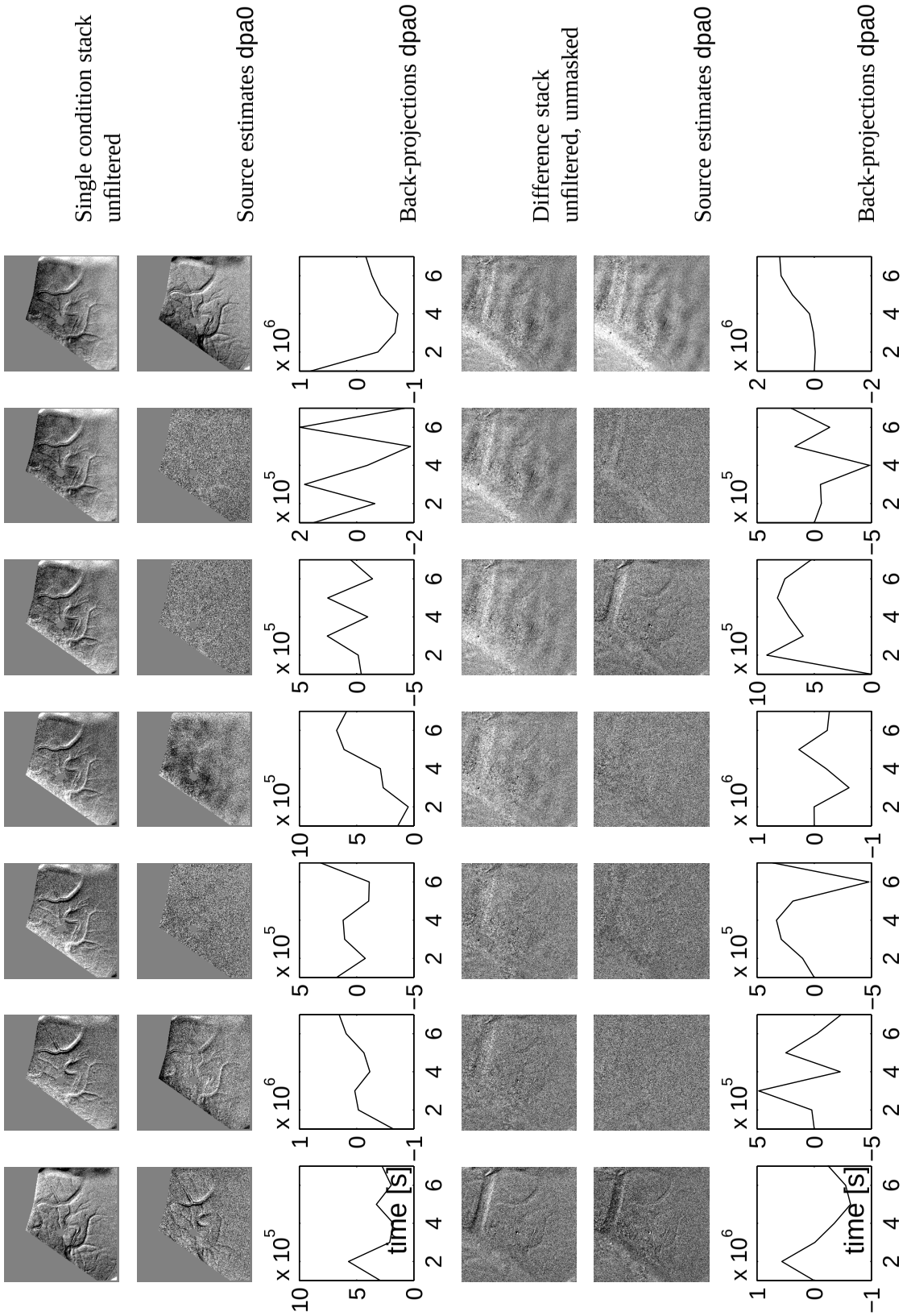


Figure 6.11: Blind Source Separation on single condition stack (top) and difference stack (bottom) for ocular dominance experiments without lowpass filtering. The difference image separation is performed for unmasked data. The algorithm used is dpa0.

Chapter 7

Discussion

In the beginning the main goal of this diploma thesis consisted in the development of a blind source separation algorithm, which would be robust against noise, and perform superior to other known algorithms on the optical imaging data set. The basis for this algorithm was the Extended Spatial Decorrelation (ESD) algorithm, presented in [MS94, SSM+99], and the result is the accelerated gradient descent algorithm. Later the algorithm called Jacobi method in this paper was published with a more noise robust sphering method in [ZM98, MPZ99]. As this algorithm is closely related to the one developed in this work, it became important to benchmark different variants of spatial decorrelation algorithms on different data sets, concerning their noise robustness.

Blind Source Separation was introduced in [HJ86]. At the moment three main directions of research can be identified in the BSS area. One is the idea presented in [MS94, SSM+99] (ESD algorithm), to use shifted correlations as additional constraints, compared with basic Principle Component Analysis. Latter itself is only able to give uncorrelated components, which can still be statistically dependent. The components found by PCA need to be rotated further to make them independent. Only if the mixing matrix is symmetrical, PCA can extract independent sources. Shifted correlation matrices can provide further information about the structure of the sources, which allows to determine the correct rotation.

Another approach is rooted in information theory. An example from this class of algorithms is the infomax algorithm, published in [BS95], which derives a learning rule for the weights of a neural network. Its goal is to make the outputs statistically independent by maximizing the information transfer of the network, thereby minimizing mutual information between the outputs. The authors give a connection between their (feed-forward) network model and the recurrent one used by Jutten and Herault, by providing a formula translating the network weights from one model into the other. Also a comparison of Jutten and Herault's network model with the error minimization of the ESD algorithm is given in [MS94]. Although the learning functions are quite different, the underlying network models are isomorph.

Another way of obtaining independent sources uses information provided by certain types of higher order moments of the data. The *infomax* algorithm uses the $\tanh$-function as contrast function, thereby exploiting all orders of statistics. [Car97] proofs *infomax* to be equivalent to a Maximum Likelihood approach. Although statistical independence involves statistics of all orders, other algorithms take into account only a restricted set of the higher order statistics. This may be motivated by the fact, that the value of cumulants diminishes the more, the higher the order is. Furthermore, the number of elements in the higher order cumulant tensors rises rapidly, posing the

question of how they can be estimated using a finite data set. Well known algorithms using fourth order cumulants as contrast function are [Com94, CS93]. In [HO97] a fast fixed point algorithm is presented, which uses information about the kurtosis (a measure for the peakedness of a distribution), the diagonal elements of the fourth order cumulant tensor. Even though these algorithms do not use all orders of statistics, they have been shown to generally work well.

One algorithm of each of these groups was evaluated in [SSM+99] on the same optical imaging data set which was used in this thesis. Although the other methods have a strong theoretical background in information theory, the ESD algorithm seemed to be best adapted to the optical imaging data set by using its broad autocorrelation function, which is ignored by the other algorithms. Together with the idea to use multiple shifts this provided a good starting point for improving existing algorithms to achieve better noise robustness. For algorithms of the ICA area, many of the theoretical results about convergence are only valid in the absence of noise. Furthermore, in [SSM+99] the two ICA algorithms performed worse on noisy, but spatially smooth, data than did the ESD algorithm.

A comparison of the algorithms related to ESD, provided in this thesis, shows that, although it is the slowest algorithm, the gradient descent method is the most flexible and noise robust one. Its advantage when compared to the original algorithm presented in [MS94] is its use of multiple shifts, which makes the selection of shifts less critical and allows to average over noise. When compared to the Jacobi method, which also uses multiple shifts, its restriction of the demixing matrix during the optimization process is more appropriate than the constraint to an orthogonal (rotation) matrix the Jacobi method imposes. This is particularly true for the use of inappropriate or unreliable sphering methods in the presence of noise.

Even though the gradient descent algorithm performed best on both data sets examined in this thesis, it has some clear limitations. First, a linear mixing process is assumed in the model of BSS. It is, on the other hand, not clear at all, how the different signal components in optical imaging are mixed in the image stacks. It could be worthwhile to pursue the idea presented in [MS94], to use several shifted correlations for the estimation of nonlinearity parameters in an extended model. The difficulty would be to create a realistic nonlinear mixing model (see discussion in [SO99]).

Second, the accelerated gradient descent algorithm still has considerable problems with convergence on complex data sets. On the single condition stacks of the optical imaging data set about four or five runs of this algorithm were necessary to obtain good separation results. The other algorithms (Jacobi method, ESD) give deterministic solutions, which do not depend on parameter initializations, except for the sphering. A minimization procedure for the cost function would be needed, which reliably finds the global optimum, without constraining the demixing matrix in the way the Jacobi method does.

Third, the computation time needed for the accelerated gradient descent, although suitable for interactive work, is still too high for applications in real time environments. It would be of much help, if a realtime BSS method could be provided, which allowed the experimenter during optical imaging experiments to extract maps from the recorded image stacks, which could be viewed online on a monitor. This would allow to better control and optimize the experimental setup, and the stimulus choices and presentations. An improvement in computation speed could be achieved by using spatial binning instead of lowpass filtering, reducing the size of the images and thereby the amount of computations needed. In addition to an improvement in computation speed, a function computing the quality of separated sources is necessary in a real-time environment, because of the stability problems of the gradient descent method. On the other hand, in comparison with other algorithms the inefficiency in computation time is less relevant when considering that all

algorithms tested in this thesis need to compute a set of correlation matrices. This step consumes at least half of the time of the Blind Source Separation runs, depending on the algorithms (for a set of approximately 50 shifts for the toy and OI data sets). A promising approach to achieve speed improvements could be a reduction of the number of shifts used in this computation. This would, on the other hand, require a more careful selection of the shifts, possibly in conjunction with a heuristic which works better than the cor heuristic.

Fourth, BSS algorithms make a number of assumptions, which could be examined closer. One of these is, that for the ESD algorithm and its extensions, the number of sources must be the same as the number of sensors. They seem to be able to deal with situations, where the number of mixtures is higher than the number of real sources; some estimated sources then contain almost only noise. The use of prior knowledge about the real sources could be useful in extending the algorithms to situations, where the number of sources is higher than the number of mixtures. The prior knowledge would provide the additional constraints necessary to obtain an unambiguous solution. Additional research should also explore the behavior of BSS algorithms for the case, that the time series or spatial auto-correlation structures of signals are very similar. In the former case, the mixing matrix would be nearly singular, and hardly invertible.

Experience shows that the choice of shifts for the multi-shift algorithms is not as critical as for the ESD algorithm; still it is recommendable to use information about the auto-correlations of the sources in selecting the shifts. Some more research could be done to determine the optimal shifts, as well for multi-shift as for single-shift algorithms. For latter the cross-correlation heuristic did not work very well on optical imaging data. For an optimization of the selected shifts an iterative process could be useful, which first gives an estimate of sources for an arbitrary choice of shifts, and then uses the structure of these sources to select new shifts. These are then used to obtain improved source estimates. Another idea is to pick small, relatively homogeneous regions of the mixtures, e.g. representing vessels, or tissue without vessels, to compute local correlation matrices. These could then be used to select shifts, which are optimized to separate sources for those regions.

Like the analysis of difference stacks the use of cocktail blanks for a normalization and signal to noise ratio improvement could provide better ocular dominance (and other) maps. Cocktail blanks are obtained by averaging images recorded from cortex which was stimulated by a presumably complete set of stimuli. The use of an “activate blank” could have the advantage that vessel artifacts, which have different blood flow and size, are canceled out better than using an “inactive blank”. The first frame, used in first frame analysis in this thesis, is an “inactive blank”, because it is recorded before stimulus presentation. The disadvantage of the use of cocktail blanks would be a problem in interpretation of the results. During first frame analysis no other stimulus condition is introduced into the image stack. The creation of cocktail blanks, on the other hand, needs some assumptions about what the complete set of stimuli for this region of the cortex is. It does also pose some problems in how to present stimuli presumably activating the whole analyzed cortex area uniformly; often cocktail blanks are created by computationally combining the recordings for several stimuli.

The experiments on the artificial and the optical imaging data sets indicated that the accelerated gradient descent algorithm is noise robust and gives good separation results. Some discrepancies between results for the artificial and the optical imaging data sets remain to be analyzed. The advantage in stability of the Jacobi algorithm when compared to the gradient descent method is stronger for the OI data set than for the artificial one. Also the bad performance of the cor heuristic is not finally clarified. The higher number of mixtures in the OI data set could be an explanation; but it could also be possible, that the assumptions about the sources not being correlated is not

completely true for the OI data set.

It still remains to apply the different methods to other data sets. One task would be to test the applicability of the gradient descend algorithm to other optical imaging experiments, e.g. Calcium imaging or fMRI. Such experiments could confirm its usefulness for practical work. In addition, artificial data sets with different autocorrelation statistics could help to explore the influence of correlation structure and number of sources on optimal shift sets, as well as on noise robustness.

An important field for future research will be the development of methods for including prior knowledge about sources and mixing process into Blind Source Separation. One point is that the mixing matrix has to be causal, i.e. the mapping signal cannot be mixed into images which are recorded before the stimulus is presented. Further knowledge which could be useful are assumptions about spatial and temporal patterns for different sources. The mapping component is relatively slow and lasts over some seconds. The pattern of the blood vessels is also known very well. Furthermore, the relative amplitudes of different signal types are known; depending on the wavelength used for recording, the mapping signal constitutes a certain percentage of the total signal (10-50% for different wavelengths for the deoxy- and the scattering signals).

In conclusion, the goal to improve the noise robustness of existing Blind Source Separation algorithms has been achieved. The accelerated gradient descent algorithm gives separation results superior or similar to those of the best evaluated existing algorithm, the Jacobi method with noise robust sphering. Also, the results for the optical imaging data set are often the best of the tested algorithms. Only if speed and stability of the algorithm are more important than separation quality, e.g. if visual inspection of results cannot be performed in real time environment, the Jacobi algorithm is preferable. Concerning the extraction of stimulus maps from optical imaging data, the ESD approach, presented in [SSM+99] and improved here, gives yet the best separation results.

Chapter 8

Acknowledgements

During my education, during my course of study at university, and during my work for this diploma thesis I got support and help by many people. First I would like to mention my parents who made my studying possible at all and always were there for me when I needed them.

Prof. Obermayer helped me very much during my scientific education, with the courses he taught at university, with the advice he gave me for my course of study and the diploma thesis, and the support he provided for my Fulbright scholarship application. I also have to thank Prof. Hommel very much for his help during this application.

Dr. Stetter has provided valuable help and many constructive discussions during my work on this thesis. His overview and explanations of optical imaging were very helpful.

Very productive were the discussions with Ingo Schiessl about the optical imaging experiments he performed, and about the Blind Source Separation work he did before I started my thesis. In conversations with him I got many valuable ideas for this work.

Many thanks go to the other members of the group I was working in during the last half year. It was fun to work and learn with Hauke, Heiner, Christian, Fabian, Thomas, Anca, Roland, Stefan, Thore, Ralf, Tatjana, Sambu, Hendrik, Robert, Michael, Jens, and especially Cornelius.

Many other friends also helped me during studying and work on this thesis, especially Stefan, Andreas, and Hendrik.

Appendix A

Derivations

A.1 Derivation of ESD Algorithm

A more detailed derivation of the standard ESD algorithm than provided in section 4.2 is given in the following.

The unshifted cross-correlation matrix is calculated as following from the given sensor signals $\mathbf{y}(\mathbf{r})$ ($\mathbf{B}$ is short for the multiplication of the sphering matrix $\mathbf{D}$ with the unknown mixing matrix $\mathbf{A}$; $\mathbf{B} = \mathbf{DA}$):

$$\begin{aligned} C_{i,j}(\mathbf{0}) &= \left\langle y'_i(\mathbf{r}) y'_j(\mathbf{r}) \right\rangle_{\mathbf{r}} \\ &= \left\langle \sum_l B_{i,l} s_l(\mathbf{r}) \times \sum_k B_{j,k} s_k(\mathbf{r}) \right\rangle_{\mathbf{r}} \\ &= \sum_l B_{i,l} B_{j,l} \lambda_l(\mathbf{0}) \end{aligned} \tag{A.1}$$

Here $\lambda_l(\mathbf{0})$ is the variance (auto-correlation of zero shift) of the source l. The last step is possible, because by assumption the cross-correlations of the sources $\langle s_l(\mathbf{r}) s_k(\mathbf{r}) \rangle_{\mathbf{r}}$ are zero for $l \neq k$. In matrix notation this is

$$\begin{aligned} \mathbf{C}(\mathbf{0}) &= \mathbf{A}\mathbf{\Lambda}(\mathbf{0})\mathbf{A}^T \quad ; \text{with} \\ \mathbf{\Lambda}(\mathbf{r}) &= \begin{pmatrix} \lambda_1(\mathbf{r}) & 0 & \cdots & 0 \\ 0 & \lambda_2(\mathbf{r}) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n(\mathbf{r}) \end{pmatrix} \end{aligned} \tag{A.2}$$

Similarly the formula for the shifted cross-correlation matrix is obtained:

$$\mathbf{C}(\Delta\mathbf{r}) = \mathbf{A}\mathbf{\Lambda}(\Delta\mathbf{r})\mathbf{A}^T \tag{A.3}$$

These equations look similar to Eigensystem problems. But they are not yet such systems, because the mixing matrix $\mathbf{A}$ is in general not orthogonal and so its transpose is not equal to its inverse. Nevertheless by substituting one equation into the other an Eigensystem can be built:

$$\begin{aligned} \mathbf{A}^{T\,-1} &= \mathbf{C}^{-1}(\Delta\mathbf{r})\mathbf{A}\mathbf{\Lambda}(\Delta\mathbf{r}) && \text{(A.4)} \\ \mathbf{\Lambda}(\mathbf{0}) &= \mathbf{A}^{-1}\mathbf{C}(\mathbf{0})\mathbf{C}^{-1}(\Delta\mathbf{r})\mathbf{A}\mathbf{\Lambda}(\Delta\mathbf{r}) && \text{(A.5)} \\ \mathbf{C}(\mathbf{0})\mathbf{C}^{-1}(\Delta\mathbf{r})\mathbf{A} &= \mathbf{A}\mathbf{\Lambda}(\mathbf{0})\mathbf{\Lambda}^{-1}(\Delta\mathbf{r}) && \text{(A.6)} \end{aligned}$$

The product $\mathbf{C}(\mathbf{0})\mathbf{C}^{-1}(\Delta\mathbf{r})$ is the matrix, of which the Eigensystem has to be computed. The last formula is equation 4.12 in section 4.2.

A.2 Derivative of Cost Function

In section 4.4.2 the cost function $E(\mathbf{W})$ given in equation 4.25 is minimized by gradient descent. The derivation of it is given in the following.[1] To implement the restriction $(\mathbf{W}^{-1})_{i,i} = 1$ for all $i = 1, \ldots, N$, an auxiliary variable was introduced: $\mathbf{T} = \mathbf{W}^{-1} - \mathbf{I}$, i.e. the main diagonal of $\mathbf{T}$ is zero. In the following $\mathbf{A}_{(i,j)}$ denotes matrix $\mathbf{A}$ without its row i and column j. $\mathbf{A}^{A}_{(i,j)}$ is an adjunct of $\mathbf{A}$, i.e. the sub-determinant of element $A_{i,j}$, multiplied by $(-1)^{i+j}$). A sub-determinant is the determinant of $\mathbf{A}$, with the ith row and the jth column removed. The derivative of the cost function

$$E(\mathbf{W}) = \sum_{\Delta\mathbf{r}} \sum_{i \neq j} \left(\left(\mathbf{W}\mathbf{C}(\Delta\mathbf{r})\mathbf{W}^T \right)_{i,j} \right)^2 \qquad \text{(A.7)}$$

with respect to $\mathbf{T}$ is given by (based on a derivation given in [Lüb97]):

$$\begin{aligned} \frac{\partial E}{\partial T_{x,y}} &= \sum_{a,b} \frac{\partial E}{\partial W_{a,b}} \frac{\partial W_{a,b}}{\partial T_{x,y}} && \text{(A.8)} \\ \frac{\partial E}{\partial W_{a,b}} &= 2 \sum_{\Delta\mathbf{r}} \sum_{i} \sum_{j \neq i} \left(\sum_{k,l} W_{i,k} W_{j,l} C_{k,l}(\Delta\mathbf{r}) \right) && \text{(A.9)} \\ &\quad \left(\sum_{k,l} \delta_{i,a}\delta_{k,b} W_{j,l} C_{k,l}(\Delta\mathbf{r}) + \delta_{j,a}\delta_{l,b} W_{i,k} C_{k,l}(\Delta\mathbf{r}) \right) \\ &= 2 \sum_{\Delta\mathbf{r}} \sum_{i \neq a} \sum_{k,l,m} W_{a,k} C_{k,l}(\Delta\mathbf{r}) W_{i,l} W_{i,m} C_{b,m}(\Delta\mathbf{r}) + \\ &\quad W_{a,k} C_{l,k}(\Delta\mathbf{r}) W_{i,l} W_{i,m} C_{m,b}(\Delta\mathbf{r}) \\ \frac{\partial W_{a,b}}{\partial T_{x,y}} &= \frac{\partial}{\partial T_{x,y}} (\mathbf{I} + \mathbf{T})^{-1}_{a,b} = \frac{\partial}{\partial T_{x,y}} (-1)^{a+b} \frac{\det\left((\mathbf{I}+\mathbf{T})_{(b,a)}\right)}{\det(\mathbf{I}+\mathbf{T})} && \text{(A.10)} \end{aligned}$$

[1] In a direct comparison it turned out that a numerical differentiation is much faster; consequently that was used in the actual simulation runs.

$$
\begin{aligned}
&= (-1)^{a+b}\frac{\det(\mathbf{I}+\mathbf{T})\frac{\partial}{\partial T_{x,y}}\det((\mathbf{I}+\mathbf{T})_{(b,a)})-\det((\mathbf{I}+\mathbf{T})_{(b,a)})\frac{\partial}{\partial T_{x,y}}\det(\mathbf{I}+\mathbf{T})}{\det(\mathbf{I}+\mathbf{T})\det(\mathbf{I}+\mathbf{T})}\\
&\overset{A.11}{=} \begin{cases} \frac{(-1)^{x+y+a+b}\det((\mathbf{I}+\mathbf{T})_{(b,a)(x,y)})}{\det(\mathbf{I}+\mathbf{T})} - W_{y,x}W_{a,b} & ; x \neq b \wedge y \neq a \\ -W_{y,x}W_{a,b} & ; \text{otherwise} \end{cases}
\end{aligned}
$$

This used the derivative of determinants, which is given here (a prime, $\cdot'$, denotes the derivative):

$$
\begin{aligned}
\frac{\partial \det(\mathbf{A})}{\partial A_{i,j}} &= \sum_k \det\begin{pmatrix} A_{1,1} & \cdots & A_{1,m} \\ \vdots & & \vdots \\ A'_{k,1} & \cdots & A'_{k,m} \\ \vdots & & \vdots \\ A_{n,1} & \cdots & A_{n,m} \end{pmatrix} = \det\begin{pmatrix} A_{1,1} & \cdots & & \cdots & A_{1,m} \\ \vdots & & & & \vdots \\ 0 & \cdots & A'_{i,j}=1 & \cdots & 0 \\ \vdots & & & & \vdots \\ A_{n,1} & \cdots & & \cdots & A_{n,m} \end{pmatrix}\\
&\overset{\text{def}}{=} \det \overline{\mathbf{A}}^{i,j}\\
&= \sum_l \overline{A}^{i,j}_{i,l}\mathbf{A}^{A}_{(i,l)}\\
&= (-1)^{i+j}\det(\mathbf{A}_{(i,j)}) \qquad \text{(A.11)}
\end{aligned}
$$

Only the derivative of one row (row i) is not equal to zero, which eliminates the first sum (because the determinant of a matrix containing a row of zeros is 0). In that row only one element is unequal to zero ($\overline{A}^{i,j}_{i,j}$, which is 1), eliminating the second sum.

Bibliography

[Ama96] S. Amari. Neural learning in structured parameter spaces - natural riemannian gradient. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, 1996.

[ASGA96] A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen. Dynamics of ongoing activity: explanation of the large variability in evoked responses. *Science*, 273:1868–1871, 1996.

[BG96] T. Bonhoeffer and A. Grinvald. Optical imaging based on intrinsic signals: The methodology. In A. Toga and J. C. Maziotta, editors, *Brain mapping: The methods*, pages 55–97, San Diego, CA, 1996. Academic Press, Inc.

[BGBM93] A. Bunse-Gerstner, R. Byers, and V. Mehrmann. Numerical methods for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl.*, 14:927–949, 1993.

[BS95] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7:1129–1159, 1995.

[Car97] J. F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Lett.*, 1997.

[CB94] B. Chapman and T. Bonhoeffer. Chronical optical imagaing of the development of orientation domains in ferret area 17. *Soc. Neurosci. Abstr.*, 20:214, 1994.

[Com94] P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36:287–314, 1994.

[CS93] J. F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings-F*, 140(6):362–370, Dec 1993.

[CS96] J. F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl.*, 17:161, 1996.

[FLTG90] R. D. Frostig, E. Lieke, D. Y. Ts'o, and A. Grinvald. Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by in-vivo high-resolution optical imaging of intrinsic signals. *Proc. Natl. Acad. Sci. U.S.A.*, 87:6082–6086, 1990.

[GB96] I. Gödecke and T. Bonhoeffer. Development of identical orientation maps for two eyes without common visual experience. *Nature*, 379:251–254, 1996.

[Gir97] M. Girolami. *Self-organizing Artificial Neural Networks for Signal Separation*. PhD thesis, Department of Computing and Information Systems, Paisley University, Scotland, 1997.

[GLF+86] A. Grinvald, E. Lieke, R. D. Frostig, C. D. Gilbert, and T. N. Wiesel. Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature*, 324:361–364, 1986.

[HJ86] J. Herault and C. Jutten. Space or time adaptive signal processing by neural network models. In J. S. Denker, editor, *Neural Networks for Computing: AIP Conference Proceedings*, volume 151. American Institute for Physics, New York, 1986.

[HK49] D. K. Hill and R. D. Keynes. Opacity changes in stimulated nerve. *J. Physiol*, 108:278–281, 1949.

[HO97] A. Hyvärinen and E. Oja. A fast fixed point algorithm for independent component analysis. *Neural Comput.*, 9:1483–1492, 1997.

[KB94] D. S. Kim and T. Bonhoeffer. Reverse occlusion leads to a precise restoration of orientation preference maps in visual cortex. *Nature*, 370:370–372, 1994.

[KO99] B.-U. Koehler and R. Orglmeister. Independent component analysis using autoregressive models. In J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, *Proceedings of the ICA99 workshop*, volume 1, pages 359–363, 1999.

[Lüb97] H. Lübben. Independent component analysis, Quellenseparation von Zeitreihen. Master's thesis, Technical University of Berlin, Germany, Department of Computer Science, Oct 1997. Language: German.

[LGBS99] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski. A unifying information-theoretic framework for independent component analysis. International Journal of Computers and Mathematics with Applications (in press), 1999.

[MAZ+96] J. E. W. Mayhew, S. Askew, Y. Zheng, J. Porril, G. W. M. Westby, P. Redgraves, D. M. Rector, and R. M. Harper. Cerebral vasomotion: 0.1hz oscillation in reflected light imaging of neural activity. *NeuroImage*, 4:183–193, 1996.

[MKDF93] S. A. Masino, M. C. Kwon, Y. Dory, and R. D. Frostig. Structure-function relationships examined in rat barrel cortex using intrinsic signal optical imaging through the skull. *Soc. Neurosci. Abstr.*, 19:1705, 1993.

[MPZ99] K.-R. Müller, P. Philips, and A. Ziehe. JadeTD: Combining higher-order statistics and temporal information for Blind Source Separation (with noise). In J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, *Proceedings of the 1. ICA99 Workshop, Aussois*, volume 1, pages 87–92, 1999.

[MS94] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3637, 1994.

[Nik93] C. L. Nikias. *Higher Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[Oja92] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927–935, 1992.

[Oja97] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17:25–45, 1997.

[PFTV88] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[RS90] A. S. Rojer and E. L. Schwartz. Cat and monkey cortical columnar pattern modeled by bandpass-filtered 2D white noise. *Biol. Cybern.*, 62:381–391, 1990.

[Rüg96] S. M. Rüger. Stable dynamic parameter adaptation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems.*, volume 8, pages 225–231. MIT Press Cambridge, MA, 1996.

[SO99] M. Stetter and K. Obermayer. Simulation of scanning laser techniques for optical imaging of blood-related intrinsic signals. *J. Opt. Soc. Am. A*, 16:in press, 1999.

[SOM+97] M. Stetter, T. Otto, T. Mueller, F. Sengpiel, M. Huebener, T. Bonhoeffer, and K. Obermayer. Temporal and spatial analysis of intrinsic signals from cat visual cortex. *Soc. Neurosci. Abstr.*, 23:455, 1997.

[SSM+99] I. Schießl, M. Stetter, J. E. W. Mayhew, S. Askew, N. McLoughlin, J. B. Levitt, J. S. Lund, and K. Obermayer. Blind separation of spatial signal patterns from optical imaging records. In J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, *Proceedings of the ICA99 workshop*, volume 1, pages 179–184, 1999.

[TRS93] D. Y. Ts'o, A. W. Roe, and J. Shey. Functional connectivity within v1 and v2: Patterns and dynamics. *Soc. Neurosci. Abstr.*, 19:1490, 1993.

[WTT94] G. Wang, K. Tanaka, and M. Tanifuji. Optical imaging of functional organization in macaque inferotemporal cortex. *Soc. Neurosci. Abstr.*, 20:316, 1994.

[ZM98] A. Ziehe and K.-R. Müller. TDSEP - an efficient algorithm for blind separation using time structure. In *Artificial Neural Networks - ICANN '98*, pages 675–680. Springer Berlin, 1998.